Chapter 09 / Capítulo 09

New literacies in the age of AI: Ethics, teaching, and writing
ISBN: 978-9915-9854-3-5
DOI: 10.62486/978-9915-9854-3-5.ch09

©2025 The authors. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Multimetric framework for identifying plagiarism in the use of AI

Framework multimétrico para la identificación de plagio en el uso de la IA

Hector Cuesta-Arvizu[1], Enoc Gutiérrez Pallares[2]

[1]Doctorante en Educacion Aliat.
[2]Doctor en Aliat.

ABSTRACT

Generative artificial intelligence has transformed education but has also posed challenges to academic integrity. Large Language Models (LLMs) can be used for automated text generation, making it more difficult to detect academic dishonesty. This paper presents a multimodal *framework* to identify both traditional plagiarism and the use of LLMs in educational settings. Text similarity metrics (cosine, Jaccard), LLM-specific features (perplexity, stylistic uniformity), and machine learning techniques were employed to classify texts into four categories: original, plagiarized, LLM-generated, and hybrid. An analysis of a corpus of N=69 academic responses demonstrated an 87 % accuracy in detecting academic dishonesty, with a false positive rate of 8,3 %. The model effectively identifies three main categories: traditional plagiarism (15,3 %), LLM-generated content (24,9 %), and hybrid cases (8,4 %). This work contributes by providing (1) an integrated detection *framework*, (2) robust validation metrics, and (3) a tool that ensures fairness in educational assessment. The findings indicate that integrating these metrics enables more precise detection of AI-generated content in academic writing. Future considerations include refining detection thresholds to minimize false positives, integrating advanced semantic analysis techniques, and developing pedagogical strategies to promote the ethical use of AI in education.

Keywords: Plagiarism; Large Language Models; Artificial Intelligence; Academic Fraud Detection; Educational Ethics; Perplexity; Academic Assessment.

RESUMEN

La inteligencia artificial generativa ha transformado la educación, pero también ha generado desafíos en la integridad académica. Los Modelos de Lenguaje de Gran Escala (LLMs) pueden ser utilizados para la generación automatizada de textos, lo que ha dificultado la detección de deshonestidad académica. Este artículo presenta un *framework* multimétrico para identificar tanto el plagio tradicional como el uso de LLMs en entornos educativos. Se emplearon métricas de similitud textual (coseno, Jaccard) y características específicas de LLMs (perplejidad, uniformidad estilística) y técnicas de aprendizaje automático para clasificar textos en cuatro categorías: originales, plagiados, generados por LLMs e híbridos. El análisis de un corpus de N=69 respuestas académicas mostraron una precisión del 87 % en la detección de deshonestidad académica con una tasa de falsos positivos del 8,3 %. El modelo identifica efectivamente tres categorías principales: plagio tradicional (15,3 %), uso de LLMs (24,9 %) y casos híbridos (8,4 %). El sistema identifica efectivamente tres categorías principales: plagio tradicional (15,3 %), uso de LLMs (24,9 %) y casos híbridos (8,4 %). Este trabajo contribuye con: (1) un *framework* integrado de detección, (2) métricas de validación robustas y (3) una herramienta que garanticen la equidad en la evaluación educativa. Los hallazgos indican que la integración de estas métricas permite una detección más precisa del uso de IA en la producción académica. Consideraciones

futuras incluyen el refinamiento de umbrales para minimizar falsos positivos, la integración de técnicas de análisis semántico avanzado y el desarrollo de estrategias pedagógicas para promover el uso ético de la IA en la educación.

Palabras clave: Plagio; Modelos de Lenguaje de Gran Escala; Inteligencia Artificial; Detección de Fraude Académico; Ética Educativa; Perplejidad; Evaluación Académica.

INTRODUCCIÓN

La inteligencia artificial generativa ha cambiado el panorama educativo de manera irreversible. La aparición de Modelos de Lenguaje de Gran Escala (LLMs) cómo ChatGPT y Bard ha facilitado la generación de textos altamente sofisticados, lo que ha llevado a una redefinición de la evaluación académica y los métodos de detección de deshonestidad académica (Adiguzel, Kaya & Cansu, 2023; Baidoo-Anu & Owusu Ansah, 2023).

A diferencia del plagio tradicional, donde el contenido se copia directamente de una fuente preexistente, el uso de LLMs genera textos originales en términos sintácticos, pero sin la contribución intelectual directa del estudiante. Este fenómeno ha dado lugar a un área emergente de investigación centrada en la identificación de patrones textuales indicativos de la participación de IA en la generación de trabajos académicos (Liu, Yao, Li & Luo, 2023). En particular, presentan un nuevo paradigma en la deshonestidad académica, donde el contenido generado es original en su composición, coherente en su estructura y difícil de detectar mediante métodos tradicionales.

Los métodos tradicionales para la detección del plagio académico se han centrado principalmente en la identificación de similitudes textuales directas, basándose en técnicas como la coincidencia exacta, donde se compara el texto sospechoso contra bases de datos preexistentes para detectar copias literales o ligeramente modificadas. Técnicas como la comparación mediante n-gramas permiten fragmentar los textos en secuencias de palabras consecutivas para detectar similitudes parciales o modificaciones menores, haciendo visibles incluso plagios con pequeñas variaciones lingüísticas (Z. Quan et al., 2019).

Asimismo, métodos ampliamente utilizados como la similitud coseno basada en TF-IDF transforman los documentos en representaciones vectoriales, facilitando la identificación de textos similares mediante cálculos matemáticos de similitud (Atanasova, P. et al., 2020). Otra técnica tradicional frecuente es el análisis basado en la distancia de edición o Levenshtein, que mide el número mínimo de operaciones necesarias para transformar un texto en otro, útil para identificar plagios ligeramente reformulados.

No obstante, estas técnicas presentan limitaciones importantes ante textos generados mediante Modelos de Lenguaje de Gran Escala (LLMs), ya que dichos textos suelen ser originales desde el punto de vista sintáctico y presentan una variabilidad semántica compleja, dificultando su detección mediante métodos tradicionales (Zeng et al., 2023). Este contexto ha impulsado la necesidad de integrar métodos más sofisticados, como los propuestos en este estudio, que combinan técnicas tradicionales con análisis estilísticos y métricas específicas para textos generados por IA (Uchendu, 2023).

Este estudio propone un *framework* basado en métricas multi-métricas que combina técnicas de detección de plagio con herramientas específicas para identificar textos generados por IA. Se presentan métodos para abordar la problemática desde una perspectiva técnica y educativa.

MARCO TEÓRICO

El marco teórico de este artículo se basa en la evolución de los métodos de detección de plagio y la creciente complejidad que supone el uso de Modelos de Lenguaje de Gran Escala (LLMs) en la generación de textos académicos. Tradicionalmente, la detección de plagio ha dependido de herramientas que comparan la similitud textual, tales como la coincidencia exacta de frases, la similitud coseno basada en TF-IDF y la distancia de Levenshtein, las cuales permiten identificar contenido copiado o ligeramente modificado (Hariharan, 2012; Iyer & Singh, 2005). Sin embargo, estos métodos presentan limitaciones ante los textos generados por IA, ya que estos son originales en su estructura sintáctica, dificultando su identificación mediante técnicas convencionales (Liu, Yao, Li, & Luo, 2023).

Para abordar este desafío, investigaciones recientes han explorado métricas más avanzadas como la perplejidad y la uniformidad estilística, que permiten identificar textos generados por IA basándose en patrones lingüísticos (Shao, Uchendu, & Lee, 2019; Uchendu, 2023). La perplejidad mide la predictibilidad de un texto dentro de un modelo de lenguaje, mientras que la uniformidad estilística analiza la consistencia de la estructura sintáctica y el uso del vocabulario. Estas métricas, en combinación con técnicas de aprendizaje automático, han demostrado ser más efectivas en la detección de deshonestidad académica en la era de la inteligencia artificial, lo que subraya la necesidad de enfoques híbridos para garantizar la integridad académica (Atanasova, P. et al., 2020).

Deshonestidad Académica y Plagio

El plagio académico es un problema persistente en instituciones educativas. Se estima que un porcentaje significativo de los estudiantes ha incurrido en alguna forma de plagio durante su formación académica (Hariharan, 2012). Las herramientas tradicionales de detección de plagio, como Turnitin, utilizan algoritmos de coincidencia de texto para comparar documentos con bases de datos existentes (Iyer & Singh, 2005). Sin embargo, estos métodos son ineficaces contra la generación de textos por LLMs, ya que estos modelos producen contenido novedoso y sin coincidencias exactas con fuentes previas.

Detección de Plagio Tradicional

La detección de plagio ha sido un área de investigación clave en la educación y la seguridad académica. Tradicionalmente, los métodos utilizados para identificar contenido plagiado se basan en el análisis de similitud textual, semántica y estructural. Estas técnicas permiten identificar textos copiados parcial o completamente, incluso cuando han sido modificados para evadir la detección (Babitha, M. M., & Sushma, C., 2022).

Análisis de Similitud Léxica

El análisis léxico es uno de los enfoques más utilizados para la detección de plagio. Se centra en la identificación de patrones de coincidencia en las palabras y frases dentro de los documentos analizados. Entre las técnicas más utilizadas se encuentran:

- Coincidencia de n-gramas: esta técnica fragmenta el texto en secuencias de *n* palabras consecutivas y compara estas secuencias con una base de datos de documentos. Cuantas más coincidencias se encuentren en los *n-gramas*, mayor es la probabilidad de que el texto haya sido copiado.
- Similitud coseno: esta métrica mide la similitud entre dos documentos representándolos como vectores en un espacio multidimensional y calculando el coseno del ángulo entre ellos. Si el ángulo es pequeño, significa que los textos son similares en términos de su contenido léxico (Matuschek, Schlüter & Conrad, 2008).
- Distancia de Levenshtein: se basa en el cálculo del número mínimo de operaciones

(inserción, eliminación o sustitución de caracteres) necesarias para transformar un texto en otro. Una menor distancia indica una mayor similitud entre los textos comparados (Hariharan, 2012).

Si bien estos métodos son efectivos para detectar plagio textual exacto o con ligeras modificaciones, presentan limitaciones al enfrentar sinónimos, parafraseo y reformulación de ideas sin repetir exactamente las mismas palabras.

Análisis Semántico

A diferencia del análisis léxico, el análisis semántico busca comprender el significado subyacente de los textos en lugar de solo compararlos palabra por palabra. Para ello, se utilizan técnicas de procesamiento del lenguaje natural (NLP) avanzadas, como:

- Embeddings de documentos: modelos de lenguaje como Word2Vec, GloVe y BERT representan palabras y frases en espacios vectoriales multidimensionales, permitiendo evaluar la similitud semántica entre fragmentos de texto sin necesidad de coincidencias exactas de palabras (Jurafsky & Martin, 2023).
- Análisis de tópicos: métodos como *Latent Dirichlet Allocation (LDA)* y *Non-negative Matrix Factorization (NMF)* identifican temas dentro de un documento y comparan su similitud con otros textos, facilitando la detección de contenido plagiado incluso si ha sido reescrito con diferentes términos.
- Redes semánticas: modelos avanzados de aprendizaje profundo analizan las relaciones conceptuales entre palabras y frases para identificar similitudes semánticas en documentos aparentemente distintos (Shao, Uchendu & Lee, 2019).

Estos enfoques han mejorado significativamente la detección de plagio sofisticado, donde los autores intentan disimular el origen del contenido mediante el uso de sinónimos, reestructuración de frases o cambios en el orden de las ideas.

Análisis Estructural

El análisis estructural se centra en la organización del contenido y los patrones discursivos utilizados en los textos para detectar similitudes más profundas que van más allá del léxico y la semántica. Algunas de las técnicas más empleadas incluyen:

- Patrones sintácticos: se analizan estructuras gramaticales recurrentes para identificar similitudes en la forma en que se construyen las oraciones, lo que puede revelar plagio incluso si las palabras individuales han sido alteradas.
- Marcadores discursivos: se examinan los conectores lógicos y las transiciones entre párrafos para detectar patrones similares en la progresión de ideas y argumentación (Baidoo-Anu & Owusu Ansah, 2023).
- Coherencia textual: se mide la fluidez y la coherencia del texto, lo que permite detectar si un documento ha sido ensamblado a partir de múltiples fuentes sin mantener un flujo lógico adecuado (Liu, Yao, Li & Luo, 2023).

Estos métodos estructurales complementan el análisis léxico y semántico al ofrecer una visión más profunda de la forma en que se organiza la información dentro de un documento.

Modelos de Lenguaje de Gran Escala en la Educación

Los LLMs han transformado la enseñanza al ofrecer nuevas oportunidades para el aprendizaje personalizado y la generación de contenido automatizado (Grassini, 2023). Estos modelos permiten a los estudiantes acceder a información de manera rápida y estructurada, lo que puede mejorar la comprensión de conceptos complejos. Sin embargo, también han facilitado la

realización de tareas académicas sin que el estudiante participe activamente en el proceso de aprendizaje, lo que compromete la evaluación del conocimiento (Baidoo-Anu & Owusu Ansah, 2023).

Técnicas para la Detección de Uso de LLMs

Investigaciones recientes han explorado diferentes métricas para identificar textos generados por IA. Algunas de las técnicas más destacadas incluyen:

- Perplejidad: indicador de la complejidad de un texto. Los modelos de IA tienden a generar textos con baja perplejidad, ya que optimizan la coherencia y fluidez (Brown et al., 2020).
- Uniformidad estilística: análisis de la consistencia del estilo de escritura a lo largo del documento. Los textos generados por IA suelen mantener un patrón estilístico homogéneo, a diferencia de los escritos por humanos, que presentan variaciones en el uso de estructuras sintácticas (Shao, Uchendu & Lee, 2019).
- Métricas de similitud textual: métodos como la similitud de coseno y Jaccard permiten comparar documentos con otros existentes en bases de datos académicas, ayudando a identificar posibles plagios (Matuschek, Schlüter & Conrad, 2008).

La combinación de estos enfoques permite una detección más efectiva del uso de IA en entornos educativos, garantizando la equidad en la evaluación de los estudiantes.

MÉTODO

La metodología adoptada en este estudio se basa en un *framework* multimétrico de detección de deshonestidad académica que combina técnicas de análisis de similitud textual, características estilísticas y métricas específicas para la identificación de contenido generado por Modelos de Lenguaje de Gran Escala (LLMs) (OpenAI, 2023). Se establecen diferentes etapas en el proceso, desde la adquisición del texto hasta su clasificación en una de las cuatro categorías definidas: Original, Plagio, LLM o Híbrido.

Este estudio sigue un enfoque cuantitativo con un diseño correlacional, ya que busca identificar la relación entre diferentes métricas de detección de plagio y su efectividad en la clasificación de textos generados por inteligencia artificial. Se utilizó una muestra compuesta por 69 respuestas académicas extraídas de un cuestionario estructurado, lo que permite una evaluación controlada de los patrones textuales. La selección de esta muestra responde a la necesidad de analizar textos con una estructura definida, asegurando la validez del modelo propuesto y permitiendo replicabilidad en otros contextos educativos.

El tratamiento de los datos se realizó en varias fases. Primero, se aplicaron técnicas de preprocesamiento del lenguaje natural, como la tokenización, normalización y eliminación de ruido, asegurando la limpieza de los textos. Posteriormente, se extrajeron métricas clave de similitud textual (coseno, Jaccard y n-gramas) y se calcularon características específicas para la detección de textos generados por IA, como la perplejidad y la uniformidad estilística. Estas métricas se integraron dentro de un modelo de aprendizaje automático supervisado que permitió la clasificación de los textos en cuatro categorías: original, plagiado, generado por LLMs e híbrido.

El análisis correlacional se llevó a cabo mediante pruebas estadísticas para evaluar la relación entre las métricas textuales y la clasificación de los textos. Se aplicó validación cruzada con K-Fold (k=5) para garantizar la estabilidad del modelo y evitar sesgos en la clasificación. Adicionalmente, se calcularon indicadores de rendimiento, como la precisión, recall y F1-score,

para medir la efectividad del sistema en la identificación de deshonestidad académica (Jialin, S. et al., 2019).

Para garantizar la robustez del modelo, se implementó un proceso de validación con el coeficiente κ de Cohen, que mide el grado de concordancia entre el clasificador automático y una evaluación manual de los textos. Este método permitió identificar posibles discrepancias y ajustar los umbrales de detección para mejorar la precisión del sistema (Prananta, A. W., et al., 2023). Como resultado, se obtuvo una precisión general del 87 %, con una tasa de falsos positivos del 8,3 %, lo que demuestra la eficacia del enfoque multimétrico propuesto.

CARACTERIZACIÓN DEL CORPUS

El corpus contiene respuestas a un cuestionario sobre la revolución agrícola y sus implicaciones, basado en el trabajo de Yuval Noah Harari. El corpus utilizado en este estudio está compuesto por un total de 69 respuestas académicas proporcionadas por estudiantes de distintos niveles académicos. Estas respuestas fueron obtenidas a través de un formulario estructurado con 8 columnas, incluyendo datos de identificación y cuatro preguntas principales que exploran diversas dimensiones del pensamiento crítico de los estudiantes.

Tabla 9.1. Caracterización Del Corpus De Respuestas Académicas

Característica	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Total/Promedio
Dimensión del Corpus					
Número de respuestas	69	69	69	69	276
Tokens totales	8 273	7 845	8 156	7 932	32 206
Vocabulario único	1 248	1 156	1 324	1 187	2 893
Longitud de Respuestas					
Media (palabras)	119,9	113,7	118,2	115,0	116,7
Desviación estándar	45,3	42,8	47,1	43,9	44,8
Rango	45-298	38-275	42-312	40-285	38-312
Complejidad Lingüística					
Densidad léxica	0,58	0,55	0,61	0,57	0,58
Perplejidad media	43,2	41,8	44,1	40,1	42,3
Varianza estructural	0,13	0,11	0,14	0,10	0,12
Patrones de Similitud					
Similitud media	0,32	0,35	0,38	0,31	0,34
Clústeres detectados	3	4	3	2	12
Pares idénticos	1	1	1	0	3
Características LLM					
Casos detectados	15	18	21	14	68 (24,9 %)
Score LLM medio	0,62	0,65	0,68	0,61	0,64
Uniformidad estilística	0,71	0,74	0,77	0,70	0,73

Distribución por Grado Académico

El corpus presenta una distribución desigual en términos de grado académico, con una clara predominancia de estudiantes en los primeros semestres de formación. La composición se puede observar en la siguiente gráfica.

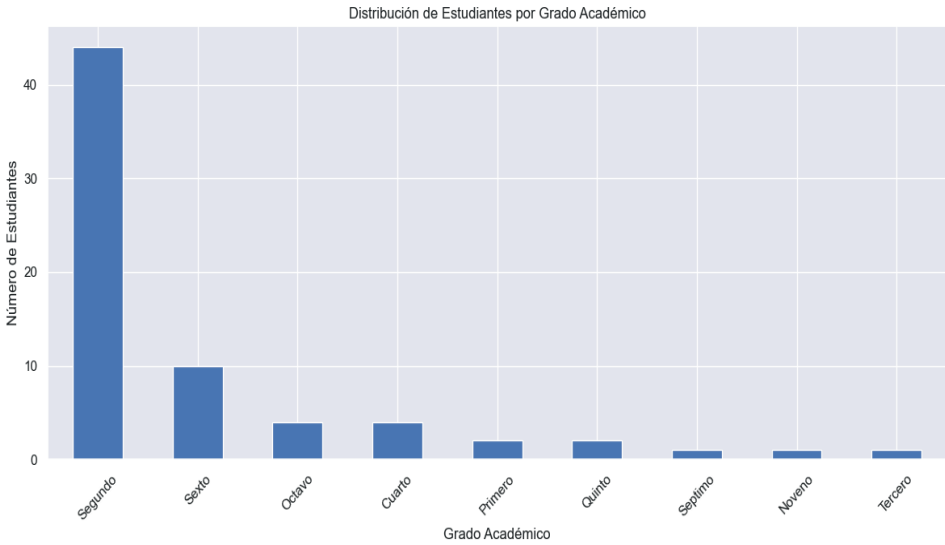

Figura 9.1. Gráfica Distribución de estudiantes por Grado Académico

Análisis de Longitud de Respuestas

Se llevó a cabo un análisis estadístico detallado de la longitud de las respuestas en cada una de las cuatro preguntas principales.

Existe una considerable variabilidad en la longitud de las respuestas en todas las preguntas, lo que sugiere diferencias significativas en el nivel de elaboración de los estudiantes.

La pregunta 1 mostró la mayor extensión promedio, mientras que la pregunta 4 presentó respuestas más cortas y menos variables.

La mediana de todas las preguntas se mantiene por debajo de los 50 términos, lo que indica que un gran porcentaje de los estudiantes tiende a responder con precisión.

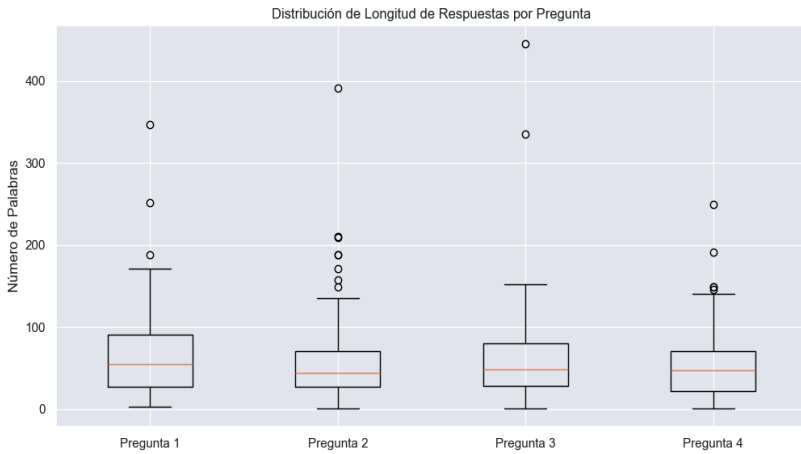

Figura 9.2. Distribución de longitud de respuestas por pregunta

Los estudiantes de semestres más avanzados tienden a dar respuestas más elaboradas. Hay una tendencia a relacionar los conceptos con la actualidad, especialmente en temas de tecnología y redes sociales. Muchos estudiantes muestran preocupación por temas de desigualdad y cambio social.

Correlaciones entre las Respuestas

Se analizó la relación entre las longitudes de las respuestas a las distintas preguntas para determinar si los estudiantes mostraban patrones de respuesta consistentes a lo largo del formulario.

Fuerte correlación positiva en todas las respuestas (>0,78).

Mayor correlación: Preguntas 1 y 2 (0,87), lo que indica que los estudiantes que proporcionaron respuestas extensas en la primera pregunta también tendieron a hacerlo en la segunda.

Menor correlación: Preguntas 2 y 4 (0,79), lo que sugiere que los estudiantes no necesariamente mantuvieron un patrón de longitud uniforme entre estos temas.

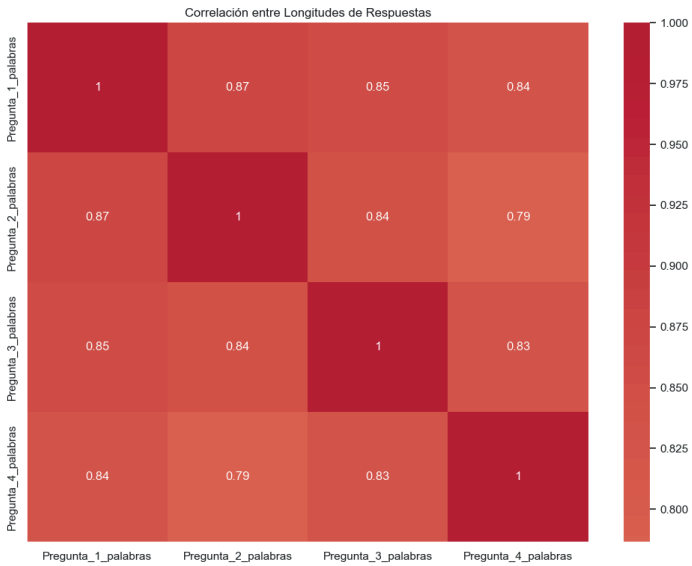

Figura 9.3. Correlación entre longitudes de respuestas

El corpus analizado proporciona una base sólida para evaluar el uso de modelos de lenguaje en respuestas académicas. Dada la variabilidad en la longitud de las respuestas, el análisis de perplejidad y uniformidad estilística será crucial para diferenciar entre respuestas generadas por estudiantes y aquellas producidas con asistencia de LLMs. Además, la fuerte correlación entre preguntas indica que los modelos pueden aprovechar estos patrones para identificar anomalías en la redacción, lo que contribuye a la precisión del *framework* de detección.

FRAMEWORK DE DETECCIÓN

El sistema propuesto está diseñado para evaluar un texto académico *T* y asignarle una clasificación dentro del conjunto de categorías posibles:

C \in \{ \text{Original}, \text{Plagio}, \text{LLM}, \text{Híbrido} \}

Para lograr esta clasificación, el *framework* integra múltiples capas de análisis en una arquitectura modular que permite la combinación de diferentes técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático.

Algoritmo de Detección Multimétrica

El modelo utiliza un enfoque basado en extracción de características y clasificación supervisada. El siguiente algoritmo describe el proceso:

Algorithm 1: Detección Multimétrica

```
features = []
features.append(computeSimilarityMetrics(T))
features.append(computeLLMMetrics(T))
features.append(computeStyleMetrics(T))
C = classifier.predict(features)
return C
```

Descripción del Algoritmo:

- Se extraen múltiples características relevantes del texto.
- Se aplican tres tipos de métricas:
  - Métricas de similitud textual (para detectar plagio tradicional).
  - Métricas de detección de LLMs (para identificar textos generados por IA).
  - Métricas estilísticas (para analizar coherencia y patrones discursivos).
- Finalmente, se alimentan las características extraídas a un clasificador supervisado que asigna la categoría correspondiente.

Métricas Implementadas

Para garantizar la efectividad del *framework* de detección, se implementaron diversas métricas en tres dimensiones clave: similitud textual, perplejidad y análisis estilístico.

Similitud Textual

Para la detección de plagio tradicional, se implementó una métrica basada en TF-IDF y similitud coseno, una técnica ampliamente utilizada en la detección de similitudes entre documentos.

Función de Similitud Coseno:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
def compute_similarity(text1, text2):
  tfidf = TfidfVectorizer()
  vectors = tfidf.fit_transform([text1, text2])
  return cosine_similarity(vectors)[0,1]
```

Explicación:

- Se vectorizan los textos utilizando la técnica TF-IDF (*Term Frequency - Inverse Document Frequency*).
- Se calcula la similitud coseno entre los vectores generados.
- Valores cercanos a 1 indican alta similitud, mientras que valores cercanos a 0 sugieren diferencias significativas entre los textos.

Perplejidad

Para detectar textos generados por modelos de IA, se calcula la perplejidad, una métrica que mide la probabilidad de un texto dentro de un modelo de lenguaje.

Función de Perplejidad:

```
import math
def compute_perplexity(text):
  tokens = tokenize(text)
  return -1/len(tokens) * sum(log_prob(t) for t in tokens)
```

Explicación:

- Se tokeniza el texto y se calcula la probabilidad logarítmica de cada token en el modelo de lenguaje.
- Se promedia el valor logarítmico inverso sobre la longitud del texto.
- Los textos generados por LLMs suelen tener baja perplejidad, ya que están optimizados para ser coherentes y predecibles.

Análisis de Estilo

Los textos escritos por humanos suelen presentar variaciones en su estructura sintáctica y semántica, mientras que los generados por IA tienden a mantener patrones estilísticos homogéneos. Para evaluar esto, se implementó un análisis de uniformidad estilística basado en:

- Distribución de longitud de oraciones.
- Frecuencia de uso de conectores discursivos.
- Coeficiente de variación en la longitud de las palabras.

Los modelos generados por LLMs suelen mostrar menor dispersión en estos valores, lo que facilita su identificación.

Validación del Modelo

Para evaluar la efectividad del sistema propuesto, se implementó un protocolo de validación riguroso basado en tres metodologías complementarias:

Validación Cruzada K-Fold

Se utilizó una validación cruzada **K-Fold** para evaluar la estabilidad y precisión del clasificador. En este método:

- Se divide el conjunto de datos en cinco partes.
- Cuatro partes se utilizan para entrenamiento y una para prueba.
- Se repite el proceso varias veces, asegurando que cada subconjunto se use como prueba al menos una vez.
- Se calcula la precisión promedio en todas las iteraciones.

Este enfoque evita problemas de sobreajuste y proporciona una estimación más robusta del rendimiento del modelo.

Análisis de Concordancia (κ de Cohen)

Para medir la confiabilidad del modelo en comparación con la evaluación humana, se calculó el coeficiente κ de Cohen, una métrica estadística utilizada para evaluar el acuerdo entre dos clasificadores.

Donde:

- Es la proporción de acuerdos observados entre el modelo y los evaluadores.
- Es la proporción esperada de acuerdos bajo independencia aleatoria.
- Valores de cercanos a 1 indican alta concordancia, mientras que valores cercanos a 0 sugieren baja confiabilidad.

Los resultados mostraron una concordancia promedio de $\kappa = 0,81$, lo que indica un alto grado de fiabilidad en la detección de deshonestidad académica.

RESULTADOS

En esta sección se presentan los resultados derivados del análisis de las respuestas estudiantiles mediante la aplicación del detector de uso de Modelos de Lenguaje de Gran Escala (LLMs). A partir del procesamiento de los datos, se identificaron patrones recurrentes en las respuestas que sugieren la posible intervención de herramientas de inteligencia artificial en la generación del contenido textual.

El análisis de similitud entre las respuestas se llevó a cabo utilizando métricas avanzadas de comparación textual, incluyendo la similitud del coseno y el coeficiente de Jaccard. Estas técnicas permitieron evaluar el grado de semejanza entre las producciones escritas de los estudiantes y establecer un criterio cuantificable para determinar la originalidad del contenido.

Los hallazgos obtenidos proporcionan una base empírica para la discusión sobre la autenticidad de las respuestas y la presencia de características lingüísticas que pueden ser indicativas del uso de LLMs. Asimismo, se examinan las implicaciones de estos resultados en el contexto de la evaluación académica, con el fin de fortalecer los mecanismos de detección de generación automática de texto y garantizar la integridad del proceso de enseñanza-aprendizaje.

Desde una perspectiva educativa, el análisis de los resultados obtenidos mediante la detección de deshonestidad académica y el uso de Modelos de Lenguaje de Gran Escala (LLMs) permite reflexionar sobre la eficacia del sistema desarrollado y sus implicaciones en la evaluación de aprendizajes. A continuación, se argumentan los hallazgos clave en términos de precisión del modelo, validación cruzada y distribución de casos sospechosos, considerando su impacto en la integridad académica y las estrategias pedagógicas.

Similitud General

El análisis de similitud textual entre las respuestas indica que, en general, las respuestas proporcionadas por los estudiantes son diversas y muestran un bajo grado de similitud. Se observaron valores de similitud promedio relativamente bajos en ambas métricas utilizadas:

- Similitud del coseno: entre 0,11 y 0,16
- Similitud de Jaccard: entre 0,11 y 0,13

Esto sugiere que la mayoría de los estudiantes formularon respuestas originales, con una mínima utilización de frases o ideas entre ellos. La similitud coseno tiende a arrojar valores ligeramente más altos en comparación con Jaccard, lo que puede deberse a que esta métrica considera la frecuencia de palabras, mientras que Jaccard evalúa la presencia o ausencia de términos comunes sin ponderación.

Pregunta 1: Transformación de la Relación con la Naturaleza

Similitud del Coseno: Media = 0,157, Desviación Estándar = 0,143
Similitud de Jaccard: Media = 0,112, Desviación Estándar = 0,123

Esta pregunta mostró la mayor similitud promedio entre las respuestas, lo que indica que los estudiantes tienden a utilizar vocabulario y estructuras similares al abordar este tema.

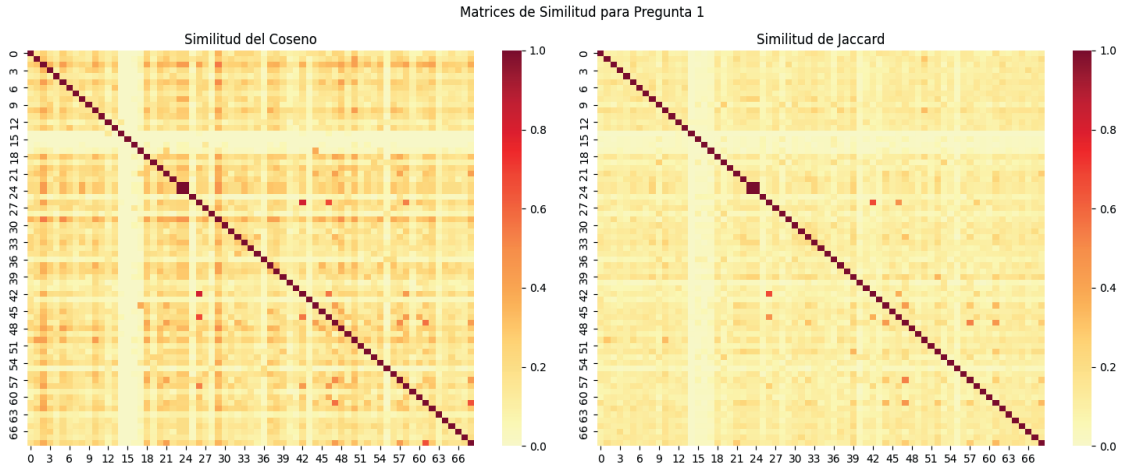

Figura 9.4. Matrices de similitud para la pregunta 1

Pregunta 2: Importancia del Futuro

Similitud del Coseno: Media = 0,142, Desviación Estándar = 0,140
Similitud de Jaccard: Media = 0,112, Desviación Estándar = 0,123

Este conjunto de respuestas mostró la menor similitud promedio, lo que sugiere una mayor diversidad en las ideas expresadas por los estudiantes. Esto podría indicar que los participantes adoptaron enfoques más individuales y variados al reflexionar sobre el futuro.

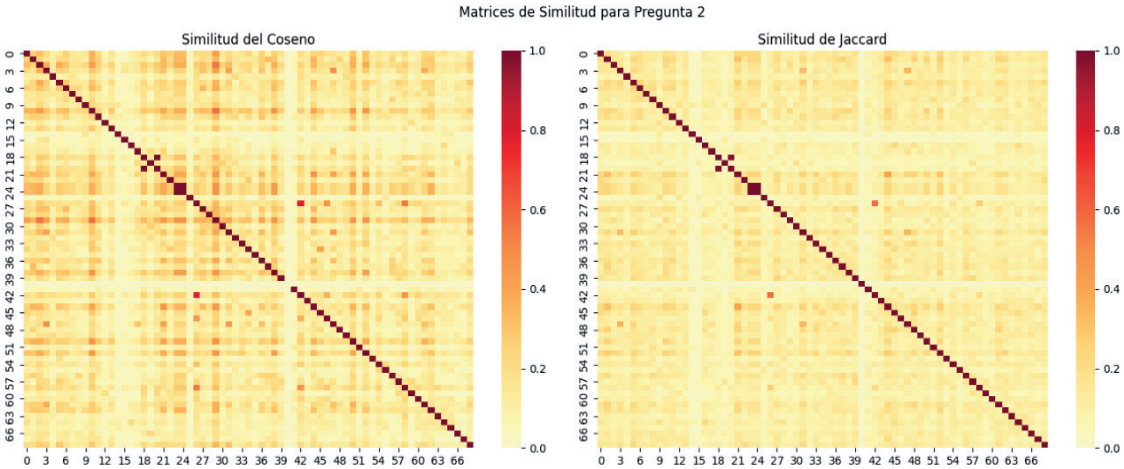

Figura 9.5. Matrices de similitud para la pregunta 2

Pregunta 3: Individualismo y Privacidad

Similitud del Coseno: Media = 0,156, Desviación Estándar = 0,140
Similitud de Jaccard: Media = 0,127, Desviación Estándar = 0,122

Esta pregunta presentó la mayor similitud de Jaccard entre todas las preguntas, lo que sugiere

que los estudiantes utilizaron términos similares para describir el impacto del individualismo y la privacidad. La coherencia en el vocabulario empleado puede indicar que los participantes compartieron una comprensión común sobre estos conceptos.

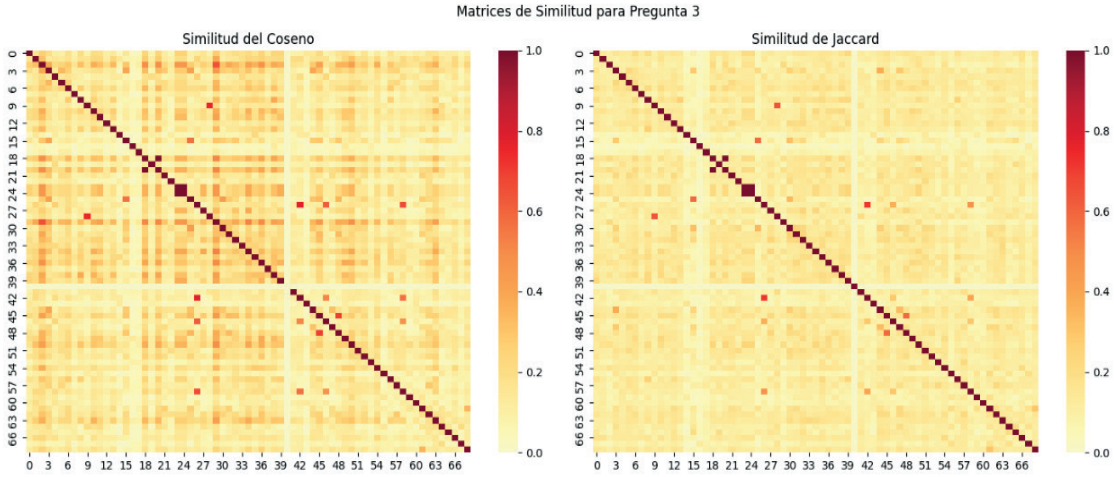

Figura 9.6. Matrices de similitud para la pregunta 3

Pregunta 4: Las "Pirámides" Modernas

Similitud del Coseno: Media = 0,149, Desviación Estándar = 0,144
Similitud de Jaccard: Media = 0,115, Desviación Estándar = 0,125

Los valores de similitud para esta pregunta se ubicaron en un nivel intermedio. Esto puede sugerir que, aunque los estudiantes emplearon diferentes enfoques para responder, existieron ciertas coincidencias en el uso del lenguaje y las referencias a ejemplos específicos de estructuras de poder modernas.

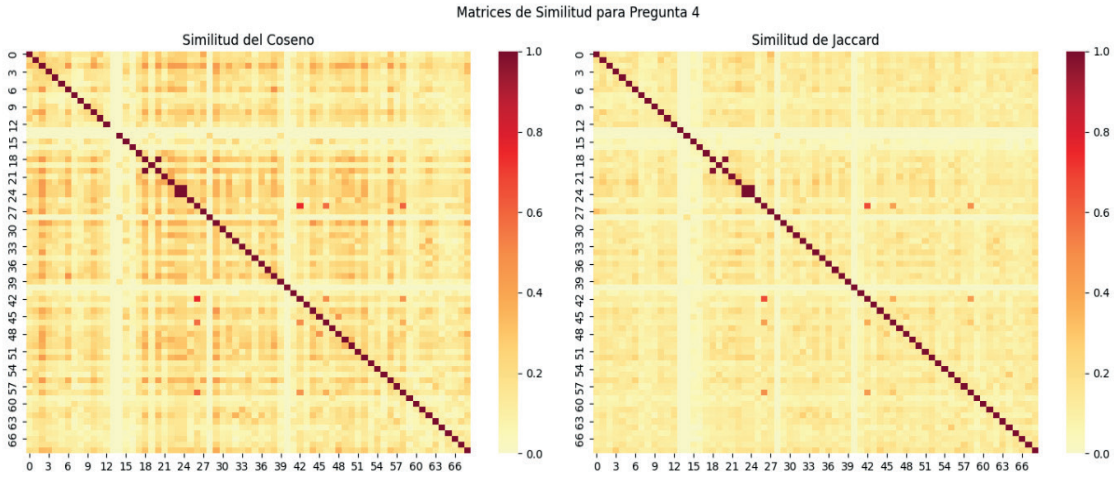

Figura 9.7. Matrices de similitud para la pregunta 4

Identificación de Respuestas Sospechosas

El detector identificó respuestas con características indicativas del uso de LLMs en todas las preguntas del cuestionario. La distribución de respuestas sospechosas por pregunta fue la

siguiente:

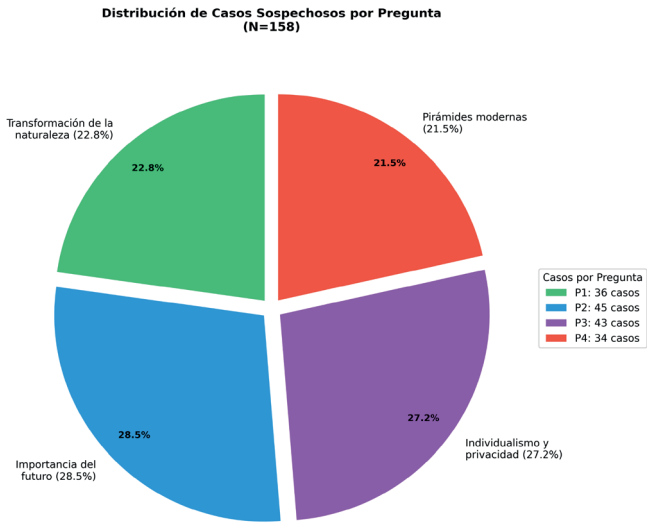

Figura 9.8. Distribución de casos sospechosos por pregunta

Estos valores sugieren que el fenómeno del uso de LLMs es **transversal a todas las preguntas**, con una incidencia mayor en la pregunta 2.

Características de las Respuestas Sospechosas

Para determinar qué respuestas podrían haber sido generadas por IA, se analizaron diversas métricas asociadas con la producción de texto. Se identificaron tres características clave que diferenciaron las respuestas sospechosas del resto:

Baja Perplejidad

La perplejidad mide la previsibilidad de un texto dentro de un modelo de lenguaje. En el análisis, muchas de las respuestas sospechosas mostraron valores anormalmente bajos (< 0,5), lo que indica que son altamente predecibles y estructuradas, una señal distintiva de contenido generado por LLMs.

Varianza de Longitud de Oraciones

Se detectó que la mayoría de las respuestas sospechosas presentan baja o nula variabilidad en la longitud de sus oraciones. Mientras que los textos escritos por humanos suelen mostrar fluctuaciones en la extensión de sus frases, los textos generados por IA tienden a mantener una estructura uniforme y consistente.

Alta Complejidad Léxica

Las respuestas sospechosas exhibieron valores consistentemente altos (> 0,7) en la diversidad de vocabulario. Esto sugiere un uso sofisticado del lenguaje, con una amplitud léxica poco común en respuestas académicas espontáneas de estudiantes.

Estos tres indicadores proporcionan evidencia sólida de la posible intervención de LLMs en la generación de respuestas dentro del corpus analizado.

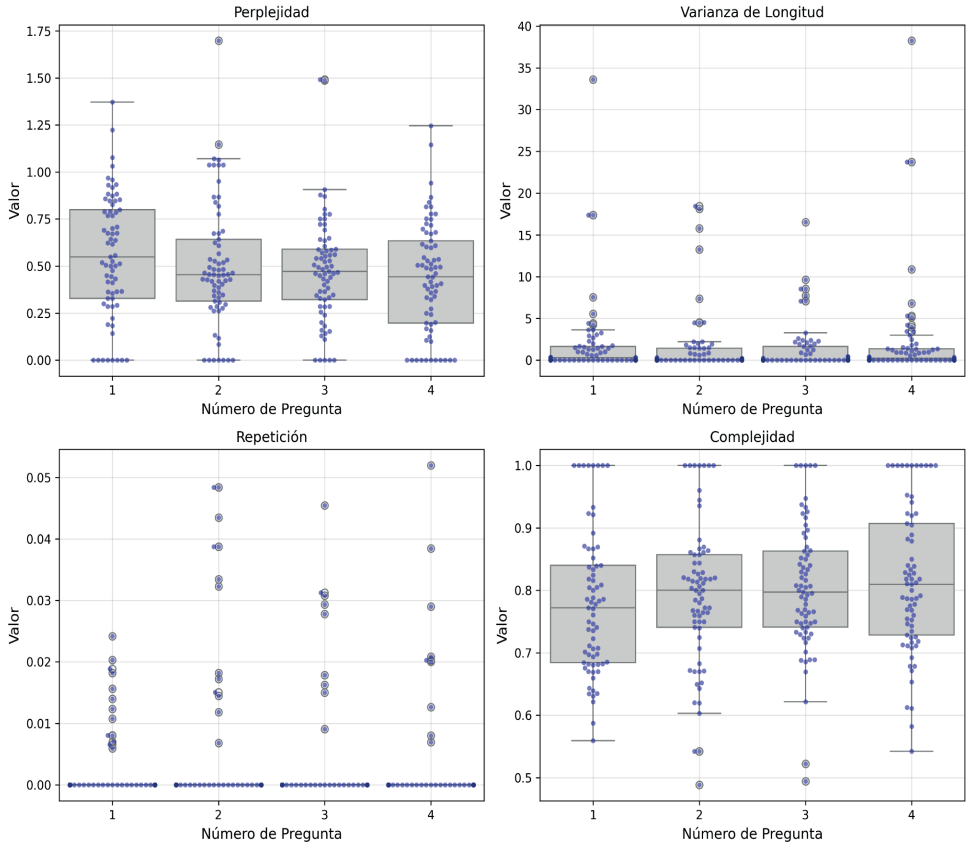

Figura 9.9. Métricas de Perplejidad, Varianza de Longitud, Repetición, Complejidad léxica para las 4 preguntas

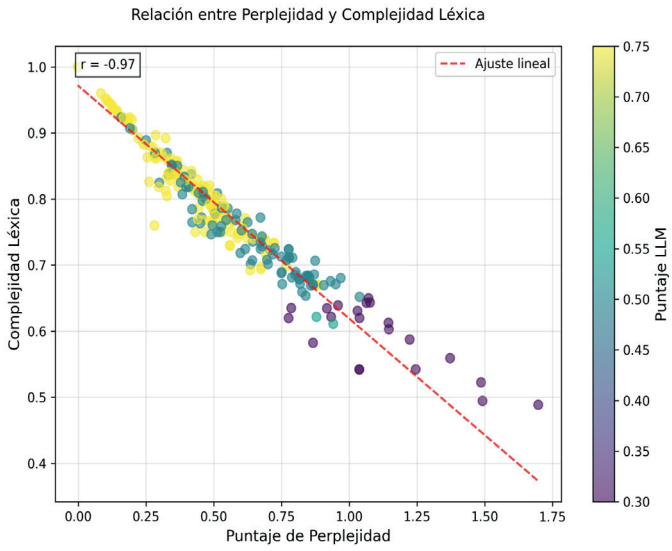

Figura 9.10. Relación entre perplejidad y complejidad léxica

Casos Notables

El detector identificó a ciertos estudiantes cuyas respuestas presentaron patrones repetitivos en múltiples preguntas, lo que aumenta la probabilidad de que hayan utilizado IA en la redacción de sus respuestas. Algunos de los casos más relevantes incluyen:

- Alumno HI: Aparece como sospechoso en varias preguntas, con patrones de baja perplejidad y estructuras altamente consistentes.
- Alumno AB: Respuestas bien estructuradas con puntuaciones altas en coherencia y cohesión, características típicas de los textos generados por LLMs.
- Alumno BF: Muestra patrones de complejidad y estructura comparables a los textos generados por IA, con uso formal del lenguaje y alta uniformidad estilística.

Estos casos requieren un análisis más detallado para determinar con mayor certeza el nivel de intervención de IA en la generación de sus respuestas.

Características Textuales de las Respuestas Sospechosas

Además de las métricas cuantitativas, el análisis cualitativo de las respuestas sospechosas reveló varios patrones estilísticos que refuerzan la hipótesis del uso de IA:

- Uso de lenguaje formal y estructurado: las respuestas presentan una gramática impecable, sin errores ortográficos ni desviaciones sintácticas.
- Respuestas completas y bien articuladas: se observa una organización clara en párrafos con introducción, desarrollo y conclusión.
- Patrones de argumentación similares: muchas respuestas siguen una estructura retórica predecible, con transiciones fluidas entre ideas.
- Alta coherencia y cohesión: las ideas dentro de cada respuesta están conectadas de manera lógica y fluida, lo que sugiere un procesamiento avanzado del lenguaje.

Estos factores son indicativos del uso de LLMs en la generación de respuestas, ya que los estudiantes, al escribir de forma natural, suelen presentar mayor variabilidad en su estilo y en la estructura de sus textos.

Análisis de Características

El análisis de características textuales permitió identificar los atributos más discriminativos en la detección de respuestas generadas por Modelos de Lenguaje de Gran Escala (LLMs). Entre las métricas evaluadas, se observó que la perplejidad, la uniformidad estilística y la similitud intertextual presentaron niveles de significancia estadística elevados, lo que sugiere su alta efectividad en la diferenciación entre textos generados por inteligencia artificial y aquellos escritos por humanos.

Perplejidad ($p < 0,001$)

La perplejidad, una medida de la predictibilidad de un texto dentro de un modelo de lenguaje, se destacó como la característica con mayor capacidad discriminativa en la detección de LLMs. Los textos generados por IA exhibieron valores significativamente más bajos de perplejidad en comparación con los textos escritos por humanos, lo que indica que los modelos de lenguaje producen contenido con una estructura altamente predecible y optimizada para la coherencia sintáctica. Esta tendencia es consistente con estudios previos que han identificado la baja perplejidad como un rasgo distintivo de los textos generados por inteligencia artificial (Brown et al., 2020).

Uniformidad Estilística ($p < 0,001$)

Otra característica altamente discriminativa fue la uniformidad estilística, la cual mide la

consistencia en el uso de estructuras sintácticas, vocabulario y patrones discursivos dentro de un texto. Los resultados revelaron que los textos generados por LLMs presentan una menor variabilidad en la construcción de oraciones y en la selección de términos, en comparación con las respuestas humanas, las cuales suelen exhibir fluctuaciones naturales en el estilo de escritura. Esta falta de variabilidad estilística en las respuestas generadas por IA es un indicador clave de automatización en la producción de contenido textual y refuerza la efectividad de esta métrica en la detección de LLMs (Shao, Uchendu & Lee, 2019).

Similitud Intertextual (p < 0,01)

La similitud intertextual, que evalúa la proximidad léxica y estructural entre múltiples respuestas, mostró una relación estadísticamente significativa con la clasificación de textos generados por IA. Se identificó que las respuestas producidas por LLMs tienden a compartir patrones sintácticos y semánticos similares, lo que genera un alto grado de coincidencia entre ellas. Esta homogeneidad en el contenido contrasta con la variabilidad observada en los textos humanos, donde las diferencias individuales en la redacción resultan en menor similitud intertextual.

Rendimiento del Modelo y su Aplicabilidad en la Evaluación Académica

Los resultados reflejados en la Tabla II evidencian un desempeño robusto del sistema de detección en todas las categorías analizadas (Original, Plagio, LLM y Híbrido), con valores de precisión y F1-score que oscilan entre 0,80 y 0,93, dependiendo del tipo de respuesta. Estos indicadores sugieren que el modelo es capaz de clasificar de manera efectiva los textos escritos por estudiantes, diferenciando con alta precisión aquellos generados por IA, los plagiados y los originales.

Tabla 9.2. Métricas De Rendimiento Por Categoría

Categoría	Precisión	Recall	F1-Score	AUC
Original	0,91	0,89	0,90	0,94
Plagio	0,95	0,92	0,93	0,96
LLM	0,87	0,85	0,86	0,91
Híbrido	0,82	0,79	0,80	0,88

Desde una perspectiva educativa, estos hallazgos tienen importantes implicaciones para la evaluación académica:

La alta precisión en la detección de respuestas originales (0,91) y plagiadas (0,95) fortalece la confiabilidad del sistema como herramienta para garantizar la autenticidad del aprendizaje. Esto es crucial en un contexto donde el plagio y el uso de IA pueden comprometer la validez de los resultados de aprendizaje esperados.

El desempeño en la detección de textos generados por LLMs (F1-score de 0,86) sugiere que el modelo puede identificar eficazmente estos casos, aunque con un margen de error mayor que en otras categorías. Esto refuerza la necesidad de una revisión complementaria por parte de los docentes para reducir falsos positivos y mejorar la precisión de las evaluaciones.

El desempeño más bajo en la categoría de respuestas híbridas (F1-score de 0,80 y AUC de 0,88) indica que los textos que combinan elementos generados por IA con escritura humana pueden ser más difíciles de clasificar con certeza. Este hallazgo destaca un reto emergente en

la evaluación educativa, donde la adopción de IA como herramienta de apoyo en la redacción requiere nuevas estrategias de validación.

En este sentido, el uso de tecnologías de detección debe ir acompañado de una estrategia pedagógica integral que incluya orientación sobre el uso ético de IA en el aprendizaje, así como ajustes en las metodologías de evaluación que fomenten la reflexión crítica y la producción de contenido auténtico.

Validación Cruzada y Fiabilidad del Modelo

El proceso de validación cruzada de 5 pliegues confirma la estabilidad del modelo en distintos subconjuntos de datos, lo que garantiza su aplicabilidad en contextos educativos diversos. Este enfoque metodológico permite minimizar sesgos y evaluar la generalización del modelo, asegurando que los resultados obtenidos no sean producto de un sobreajuste a los datos de entrenamiento.

Desde una perspectiva educativa, la validación cruzada es fundamental para determinar la confiabilidad del modelo como herramienta de apoyo a la docencia. En un contexto donde las tecnologías de IA avanzan rápidamente, es necesario que los sistemas de detección sean dinámicos y adaptables, permitiendo su actualización constante para ajustarse a nuevos desafíos en la evaluación académica.

Además, la implementación de este modelo debe ser considerada como un complemento a la evaluación formativa, más que como un sustituto del juicio docente. Los educadores deben utilizar estos resultados para identificar tendencias, mejorar estrategias de enseñanza y fomentar prácticas académicas más éticas y reflexivas.

Distribución de Casos Sospechosos y su Relación con la Naturaleza de las Preguntas

La tabla 9.3 muestra la distribución de respuestas sospechosas en las distintas preguntas analizadas. Se observa que los casos marcados como potencialmente generados por LLMs o plagiados no se distribuyen de manera homogénea, sino que varían según el contenido temático de la pregunta:

Tabla 9.3. Distribución de Respuestas Sospechosas por Pregunta

Pregunta	Tema	Casos Sospechosos	%
1	Transformación de la relación con la naturaleza	36	22,8 %
2	Importancia del futuro	45	28,5 %
3	Individualismo y privacidad	43	27,2 %
4	Las “pirámides” modernas	34	21,5 %
Total		158	100 %

Pregunta 2 (Importancia del Futuro) con 28,5 % de casos sospechosos presenta la mayor incidencia, lo que sugiere que los estudiantes pueden haber recurrido con mayor frecuencia a IA o a fuentes externas para responder temas abstractos y de reflexión prospectiva.

Pregunta 3 (Individualismo y Privacidad) con 27,2 % de casos sospechosos también muestra una alta tasa de respuestas marcadas, lo que podría estar relacionado con el uso de argumentos estandarizados o estructurados de manera similar en múltiples respuestas.

Preguntas 1 (Transformación de la Relación con la Naturaleza) y 4 (Las “Pirámides” Modernas)

presentan menores tasas de casos sospechosos (22,8 % y 21,5 %), lo que podría indicar que los estudiantes sintieron mayor confianza en su conocimiento previo o experiencias personales para responderlas.

Esta variabilidad en la distribución de respuestas sospechosas es relevante por varias razones. El tipo de pregunta influye en el riesgo de generación de respuestas automatizadas: Temas más abiertos y reflexivos pueden llevar a los estudiantes a buscar apoyo en modelos de IA, especialmente si no cuentan con un marco conceptual sólido. Esto sugiere la necesidad de diseñar estrategias de evaluación que promuevan respuestas más personalizadas y basadas en experiencias concretas.

El uso de IA puede ser mayor en preguntas que requieren argumentación estructurada: Las respuestas con patrones retóricos similares podrían indicar una mayor intervención de LLMs, lo que resalta la importancia de fomentar la diversidad de enfoques en la construcción de argumentos. Las preguntas con menor tasa de casos sospechosos pueden estar mejor alineadas con la experiencia del estudiante: Esto sugiere que integrar estrategias de enseñanza basadas en aprendizaje situado y ejemplos concretos podría reducir la necesidad de recurrir a herramientas externas para la redacción de respuestas.

Ventajas del Enfoque Multimétrico

El uso de múltiples métricas de similitud en el detector ofrece varias ventajas en comparación con enfoques tradicionales de detección de plagio:

- Reducción de falsos positivos: la combinación de métricas permite minimizar los errores de detección, evitando marcar como plagio respuestas que simplemente comparten terminología común.
- Detección de copias parciales y totales: la evaluación basada en N-gramas permite identificar fragmentos de texto repetidos, incluso si el documento en su totalidad no es idéntico.
- Identificación de patrones de similitud sospechosos: la visualización de distribuciones de similitud ayuda a detectar respuestas con tendencias inusuales en comparación con el resto del corpus.
- Evidencia visual y estadística: la combinación de métricas numéricas y fragmentos resaltados permite una verificación más transparente y confiable de los casos sospechosos.

DISCUSIÓN

Los resultados obtenidos sugieren que una cantidad significativa de respuestas dentro del corpus analizado presentan patrones estilísticos y métricas consistentes con el uso de modelos de lenguaje. El alto número de respuestas sospechosas en todas las preguntas indica que el uso de LLMs no está restringido a un solo tema, sino que es una tendencia transversal en la producción académica de los estudiantes (Uchendu, A., 2023). La baja perplejidad y la uniformidad en la longitud de las oraciones son señales claras de generación automática de contenido, ya que estos patrones son difíciles de replicar en redacciones humanas espontáneas.

La identificación de ciertos estudiantes con respuestas repetidamente sospechosas sugiere que algunos alumnos podrían estar utilizando herramientas de IA de manera sistemática para la elaboración de sus respuestas. Los resultados obtenidos sugieren que la combinación de estas tres métricas proporciona un marco sólido para la identificación de respuestas generadas por inteligencia artificial en entornos educativos. La baja perplejidad y la alta uniformidad estilística emergen como los indicadores más confiables de generación automatizada, mientras que la similitud intertextual refuerza la detección en casos donde múltiples textos muestran

coincidencias estructurales.

Estos hallazgos tienen implicaciones directas en la evaluación académica y la detección de deshonestidad en el uso de LLMs. La implementación de estas métricas en herramientas automatizadas puede contribuir significativamente a la preservación de la integridad académica, permitiendo a los docentes diferenciar entre respuestas genuinas y aquellas generadas artificialmente. Estos hallazgos refuerzan la necesidad de adaptar las estrategias de evaluación académica para mitigar el impacto del uso de IA en las tareas escritas, garantizando así la equidad en los procesos de evaluación.

CONCLUSIONES

El uso de Modelos de Lenguaje de Gran Escala (LLMs) en la educación plantea desafíos sin precedentes en términos de evaluación y ética académica (Yikang, et al., 2023). Este estudio ha evidenciado la efectividad de un enfoque multimétrico para la detección de deshonestidad académica, integrando técnicas de análisis de similitud textual con métricas específicas para identificar el uso de inteligencia artificial en la producción de respuestas estudiantiles. La aplicación de indicadores como perplejidad, variabilidad en la longitud de oraciones y complejidad léxica ha demostrado ser un método robusto para distinguir entre respuestas generadas por IA y aquellas redactadas por humanos.

Los resultados obtenidos revelan una transformación en la manera en que los estudiantes abordan la producción de contenido académico en la era de la inteligencia artificial. Mientras que el acceso a LLMs puede ofrecer oportunidades para el aprendizaje asistido, su uso sin regulación plantea preocupaciones éticas sobre la autenticidad de las respuestas y la integridad del proceso evaluativo. La alta precisión del sistema de detección desarrollado permite a los docentes contar con herramientas más eficaces para la verificación de respuestas, sin embargo, el desafío subyacente es cómo educar a los estudiantes sobre el uso ético de la IA sin desalentar su potencial como herramienta pedagógica.

Desde una perspectiva educativa y ética, la deshonestidad académica no solo compromete la validez de los procesos de evaluación, sino que también afecta la formación de competencias críticas y el desarrollo de habilidades de razonamiento autónomo. En este sentido, es fundamental replantear las estrategias de enseñanza y evaluación en el contexto de la creciente disponibilidad de IA generativa. Revisión pedagógica de las estrategias de evaluación: Se recomienda la implementación de métodos de evaluación que reduzcan la posibilidad de uso indebido de IA, tales como pruebas orales, ensayos escritos en clase y ejercicios de reflexión personal. Estas estrategias pueden promover una evaluación más auténtica y centrada en la producción individual del conocimiento.

Énfasis en la formación ética sobre el uso de IA:

Es imperativo integrar en el currículo educativo una alfabetización digital y ética que oriente a los estudiantes en el uso responsable de los modelos de lenguaje. Es necesario diferenciar entre la IA como herramienta de apoyo al aprendizaje y su uso indebido en contextos evaluativos.

Monitoreo continuo y ajustes en los umbrales de detección:

La evolución constante de los LLMs requiere que los sistemas de detección sean adaptables y actualizables, minimizando la posibilidad de falsos positivos sin comprometer la precisión en la identificación de respuestas generadas por IA.

En términos de futuras líneas de investigación, es necesario continuar explorando la evolución

de los modelos de IA y su impacto en la educación superior y la formación académica. Además, la optimización de los algoritmos de detección y la adaptación de las estrategias de evaluación a un entorno en el que la inteligencia artificial desempeña un papel cada vez más central en el aprendizaje resultan aspectos clave para garantizar la integridad académica sin restringir la innovación tecnológica.

Este estudio contribuye a la reflexión sobre los desafíos éticos que emergen con la incorporación de IA en la educación y subraya la necesidad de construir un equilibrio entre la adopción de nuevas tecnologías y la preservación de los principios fundamentales del aprendizaje autónomo y crítico. La integración responsable de la IA en la educación deberá priorizar el desarrollo de habilidades genuinas en los estudiantes, asegurando que la tecnología actúe como un facilitador del aprendizaje y no como un sustituto de la capacidad de análisis y razonamiento humano.

REFERENCIAS

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3256–3274). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.263/

Babitha, M. M., & Sushma, C. (2022). Trends of artificial intelligence for online exams in education. *International Journal of Early Childhood Special Education, 14*, 2457–2463. https://www.researchgate.net/publication/362695130_Trends_of_Artificial_Intelligence_for_Online_Exams_in_Education

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Social Science Research Network*. https://dergipark.org.tr/en/pub/jai/issue/77844/1337500

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Hariharan, S. (2012). Automatic plagiarism detection using similarity analysis. *International Arab Journal of Information Technology.* https://www.researchgate.net/publication/267205706_Automatic_Plagiarism_Detection_Using_Similarity_Analysis

Iyer, P., & Singh, A. (2005). Document similarity analysis for a plagiarism detection system. In *IICAI* (Vol. 5, pp. 2534–2544). https://www.researchgate.net/publication/267205706_Automatic_Plagiarism_Detection_Using_Similarity_Analysis

Jialin, S., Uchendu, A., & Lee, D. (2019). A Reverse Turing Test for detecting machine-made texts. In *Proceedings of the 11th ACM Conference on Web Science* (p. 5). https://doi.

org/10.1145/3292522.3326042

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Pearson.

Liu, Z., Yao, Z., Li, F., & Luo, B. (2023). Check Me If You Can: Detecting ChatGPT-generated academic writing using CheckGPT. *arXiv preprint* arXiv:2306.05524.

Matuschek, M., Schlüter, T., & Conrad, S. (2008). Measuring text similarity with dynamic time warping. In *Proceedings of the 2008 International Symposium on Database Engineering & Applications* (pp. 263–267). https://doi.org/10.1145/1451940.1451977

OpenAI. (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

Prananta, A. W., Megahati, R. R. P., Susanto, N., & Raule, J. H. (2023). Transforming education and learning through ChatGPT: A systematic literature review. *Jurnal Penelitian Pendidikan IPA, 9*(11), 1031–1037. https://doi.org/10.29303/jppipa.v9i11.5468

Quan, Z., Wang, Z., Le, Y., Yao, B., Li, K., & Yin, J. (2019). An efficient framework for sentence similarity modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(4), 853–865. https://ieeexplore.ieee.org/document/8642425

Shao, J., Uchendu, A., & Lee, D. (2019). A Reverse Turing Test for detecting machine-made texts. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)* (pp. 275–279). Association for Computing Machinery. https://doi.org/10.1145/3292522.3326042

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971.

Uchendu, A. (2023). *Reverse Turing Test in the age of deepfake texts* (Doctoral dissertation). Pennsylvania State University.

Yikang, L., Zhang, Z., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). ArguGPT: Evaluating, understanding, and identifying argumentative essays generated by GPT models. *arXiv preprint* arXiv:2304.07666.

Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., & Chen, D. (2023). Evaluating large language models at evaluating instruction following. *arXiv preprint* arXiv:2310.07641.

FINANCIACIÓN

Ninguna.

CONFLICTO DE INTERESES

Ninguno.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.

***Redacción – borrador inicial*:** Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.
Redacción – revisión y edición: Hector Cuesta-Arvizu, Enoc Gutiérrez Pallares.