Detection of AI-generated text using ensemble classifiers and stylometric feature extraction

César Espin-Riofrio, MSc.[1], Joel Alejandro Barba-Salazar, MGs.[1], Verónica Mendoza Morán, MSc.[1], Oswaldo Vergara-Bello, MSc.[1], Johanna Zumba Gamboa, MGs.[1], Josthin Ayon-Castillo, Ing.[1], Guillermo Zevallos-Escalante, Ing.[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, joel.barbas@ug.edu.ec, veronica.mendozam@ug.edu.ec, oswaldo.vergarab@ug.edu.ec, johanna.zumbag@ug.edu.ec, jhostin.ayonc@ug.edu.ec, guillermo.zevallose@ug.edu.ec

Abstract– ***The automatic generation of content has transformed the way information is produced and consumed, but it has also posed significant challenges in ensuring its authenticity and reliability, particularly in sectors such as education and media. Differentiating between automatically generated texts and those written by humans is crucial to prevent the spread of misinformation and ensure transparency in the use of these technologies. In this context, this paper proposes an effective approach based on traditional classification models combined with ensemble techniques and advanced Natural Language Processing (NLP) methods, using textual features such as phraseological measures, TF-IDF with n-grams, and perplexity to capture distinctive patterns. The methodology was evaluated on datasets from the COOLING 2025 workshop, including corpora in English, Arabic, and multilingual datasets, covering different sizes and complexities. The Stacking Classifier model achieved an F1-macro of 0.9273 on the large English corpus and 0.9131 on the multilingual corpus, demonstrating its effectiveness in diverse scenarios. Additionally, Logistic Regression and XGBoost achieved perfect performance on smaller and more homogeneous datasets in English and Arabic, respectively. These results highlight the robustness of the proposed approach, which combines key textual features with robust models, offering an effective tool to tackle the challenges of automatic content generation in multilingual and complex contexts.***

Keywords- ***Generated text, Human or Machine, Perplexity, Phraseological features, Natural Language Processing.***

Detección de texto generado por IA mediante ensamblado de clasificadores y extracción de características estilométricas

César Espin-Riofrio, MSc.[1], Joel Alejandro Barba-Salazar, MGs.[1], Verónica Mendoza Morán, MSc.[1], Oswaldo Vergara-Bello, MSc.[1], Johanna Zumba Gamboa, MGs.[1], Josthin Ayon-Castillo, Ing.[1], Guillermo Zevallos-Escalante, Ing.[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, joel.barbas@ug.edu.ec, veronica.mendozam@ug.edu.ec, oswaldo.vergarab@ug.edu.ec, johanna.zumbag@ug.edu.ec, jhostin.ayonc@ug.edu.ec, guillermo.zevallose@ug.edu.ec

Resumen* — La generación automática de contenido ha transformado la manera en que se produce y consume información, pero también ha planteado serios desafíos para garantizar su autenticidad y confiabilidad, especialmente en sectores como la educación y los medios de comunicación. Diferenciar entre textos generados automáticamente y escritos por humanos es crucial para evitar la propagación de desinformación y garantizar la transparencia en el uso de estas tecnologías. En este contexto, el presente trabajo propone un enfoque eficaz basado en modelos de clasificación tradicionales junto con técnicas de ensamblado y técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN), utilizando características textuales como medidas fraseológicas, TF-IDF con n-gramas y perplejidad para capturar patrones distintivos. La metodología fue evaluada en datasets del workshop COOLING 2025, que incluyeron corpus en inglés, árabe y multilingüe, abarcando diferentes niveles de tamaño y complejidad. El modelo Stacking Classifier alcanzó un F1-macro de 0.9273 en el corpus en inglés de gran tamaño y 0.9131 en el corpus multilingüe, demostrando su eficacia en escenarios diversos. Además, Logistic Regression y XGBoost lograron un desempeño perfecto en datasets más pequeños y homogéneos en inglés y árabe, respectivamente. Estos resultados destacan la solidez del enfoque propuesto, que combina características textuales clave con modelos robustos, ofreciendo una herramienta efectiva para abordar los retos de la generación automática de contenido en contextos multilingües y complejos.

Palabras clave*— Texto generado, Humano o máquina, Perplejidad, Características fraseológicas, Procesamiento de Lenguaje Natural.

I. Introducción

El avance acelerado de los modelos generativos de texto ha transformado profundamente la producción de contenido digital, generando una problemática con impactos significativos en múltiples sectores. La creciente sofisticación de modelos como GPT-3 [1], LaMDA [2] y LLaMA [3] ha logrado que los textos generados automáticamente se asemejen notablemente al contenido escrito por humanos, dificultando su identificación y autenticidad. Este fenómeno no solo afecta la confianza en la información, sino que también plantea desafíos éticos, sociales y prácticos en ámbitos como la educación, los medios de comunicación y el sector empresarial.

La rápida proliferación y accesibilidad de herramientas avanzadas de generación de texto han superado la capacidad de los métodos de detección existentes, generando un vacío en regulación y soluciones efectivas. Además, la falta de enfoques adaptados a distintos idiomas y su uso creciente en contextos educativos y profesionales agravan el problema.

Las implicaciones de estas tecnologías son preocupantes. En el ámbito académico, facilitan el plagio y comprometen la calidad educativa, mientras que en los medios de comunicación contribuyen a la desinformación y la polarización social. Además, la dependencia de estas herramientas en entornos profesionales podría degradar la calidad de los procesos creativos y analíticos, limitando el desarrollo de habilidades críticas [4].

Este escenario evidencia tanto la dificultad de diferenciar entre textos generados por IA y escritos por humanos como el uso indebido de herramientas creadas con otros fines. La creciente capacidad de estos modelos para imitar el lenguaje ha superado los métodos tradicionales de detección, amplificando los riesgos éticos, sociales y culturales.

Un estudio relevante es el de [5], quienes propusieron un modelo basado en embeddings de tokens iniciales de doce capas ocultas de BERT, logrando una alta precisión en la detección de textos generados automáticamente. Estos embeddings capturan características sintácticas y semánticas complejas, permitiendo diferenciar patrones propios de los textos humanos frente a los generados.

Asimismo, [6] desarrollaron un modelo centrado en la detección de errores léxico-sintácticos en textos en español mediante el Índice de Densidad de Errores (IDE). Este índice permitió una evaluación cuantitativa de la calidad sintáctica, superando la evaluación manual en términos de precisión y eficiencia. Este enfoque ha sido destacado por su aplicabilidad en contextos educativos, ayudando a identificar problemas como errores de concordancia y uso incorrecto de preposiciones, los cuales suelen ser más frecuentes en textos generados automáticamente.

Por otro lado, [7] exploraron el uso de técnicas tradicionales de PLN, como el TF-IDF[8], combinado con clasificadores avanzados como Random Forest (RF)[9] y XGBoost (XGB) [10], para la detección de patrones distintivos en textos generados por IA. Este método se enfocó en la identificación de términos sobreutilizados y patrones repetitivos, logrando mejorar la precisión en la clasificación y reduciendo falsos positivos, lo que refuerza la importancia de integrar técnicas tradicionales con enfoques modernos.

Tambien [11] implementaron un sistema basado en soft-voting equilibrado, combinando predicciones de múltiples clasificadores para mejorar la detección de textos generados. Este enfoque demostró ser efectivo en el desafío SemEval-2024[1].

Además de esto, [12] implementaron un esquema de stacking basado en modelos Transformers, integrando las capacidades individuales de diferentes arquitecturas para aumentar la precisión en la clasificación de textos. Este enfoque alcanzó una precisión del 95.55% en el desafío ALTA 2023[2].

Otro estudio relevante es el de [13], quien combinó análisis de sentimiento con un modelo de Random Forest para clasificar textos generados por modelos de lenguaje grande frente a textos humanos. Este enfoque permitió capturar características específicas del tono y estilo lingüístico, logrando una precisión general del 84.14% y una tasa F1 de 0.841, evidenciando la eficacia de los algoritmos supervisados en la detección de diferencias estilísticas.

En un contexto complementario, [14] integraron características fraseológicas y léxicas en un estudio sobre clasificación de textos en español. Utilizando modelos supervisados como Random Forest y Gradient Boosting, este trabajo destacó el uso de métricas como longitud media de palabras, diversidad léxica y características relacionadas con la estructura de oraciones y párrafos. Los resultados reflejaron mejoras significativas en tareas automatizadas de clasificación, consolidando el análisis fraseológico y léxico como herramientas clave en este campo.

De forma paralela, la métrica de perplejidad también ha sido explorada como un método para detectar texto generado automáticamente. [15] analizaron su aplicabilidad y limitaciones, concluyendo que, aunque es útil para capturar patrones de coherencia textual, su eficacia se ve comprometida en presencia de técnicas avanzadas como el parafraseo recursivo.

Esta investigación busca mejorar la detección de textos generados por IA mediante un sistema basado en clasificación y ensamblado, capaz de diferenciar contenido humano y automatizado. Enfocado en datasets en inglés, árabe y multilingües, el proyecto responde a la necesidad de herramientas adaptables y eficientes a nivel global. Su impacto va más allá del ámbito académico y profesional, fortaleciendo la confianza en la información digital y mitigando los riesgos de desinformación.

II. MÉTODO

La investigación emplea un enfoque experimental, diseñado para analizar y desarrollar métodos de detección de texto generado automáticamente. Se basa en el uso de modelos de clasificación y ensamblado, con un diseño centrado en optimizar la precisión y generalización de los resultados.

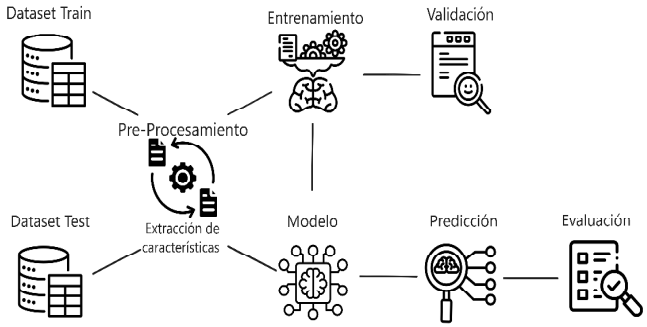

Fig. 1 Esquema del proceso realizado

El modelo propuesto se estructura en un flujo de trabajo compuesto por etapas clave que abarcan desde la preparación inicial de los datos hasta la clasificación y evaluación final. Los datasets de entrenamiento y prueba pasan por el preprocesamiento, donde se prepara la información eliminando elementos irrelevantes y estructurando los textos para su análisis; extracción de características, donde se generan representaciones relevantes mediante técnicas avanzadas; y entrenamiento, donde el modelo aprende patrones distintivos utilizando algoritmos de clasificación y ensamblado.

Tras el entrenamiento, el modelo es evaluado en la etapa de validación, utilizando métricas clave para garantizar su capacidad predictiva. Finalmente, en la fase de predicción, el modelo clasifica nuevos textos, cuyos resultados son analizados en la evaluación final para asegurar el cumplimiento de los objetivos propuestos.

A. Datasets utilizados

Los datasets utilizados en esta investigación provienen del workshop on Detecting AI Generated Content at Coling 2025[3], un evento especializado que se centra en la identificación de textos generados por modelos de inteligencia artificial. Este taller incluyó tres tareas principales, de las cuales se seleccionaron dos datasets pertenecientes a la Tarea 1, correspondientes a textos en inglés y textos multilingües, así como dos datasets adicionales de la Tarea 2, enfocados en textos en inglés y árabe. Estas colecciones fueron diseñadas para cubrir una amplia variedad de dominios textuales como

[1] https://semeval.github.io/SemEval2024/tasks
[2] https://alta2023.netlify.app/
[3] https://genai-content-detection.gitlab.io/sharedtasks

Reddit, open_qa, outfox, Arxiv y Wikipedia y contextos lingüísticos. En cuanto a la fuente de textos de generación de forma automática podemos encontrar populares modelos como GPT-3.5, GPT-3.5-Turbo, GPT-4o, GPT-4o-mini, GPT-4, Gemini-1.5, Phi-3.5-mini, Claude-3.5, Claude, Bloom. LLaMA y LLaMA-3.1 (8B).

El dataset en inglés [16] correspondiente a la tarea 1, en Tabla 1 podemos apreciar la cantidad de textos generados en comparación de textos escritos por humanos.

TABLA 1
DISTRIBUCIÒN DEL DATASET EN INGLÈS, TAREA 1

Tipo	Entrenamiento	Validación
Humano	228,992	98,328
Generado	381,845	163,430
Total	610,767	261,758

En la figura 2, podemos observar un ejemplo del dataset antes mencionado.

	id	source	sub_source	lang	model	label	text
0	5cd5f0	m4gt	peerread	en	gpt4	1	The paper titled "A Transition-Based Directed ...
1	ıc7ba0	mage	wp	en	text-davinci-003	1	(Apologies for two submissions, but need to wr...
2	i845ca	mage	cmv	en	7B	1	WARNING: WALL OF TEXT!!! I also jump from topi...
3	l464b9	m4gt	outfox	en	cohere	1	Emotion recognition through facial feedback ha...
4	:5b55a	mage	eli5	en	gpt-3.5-turbo	1	Several things. 1. The cooling effect of air c...

Fig. 2 Muestra del dataset inglés, tarea 1

El dataset multilingüe [17] incluye textos en chino, inglés, italiano, árabe, alemán, ruso, búlgaro y urdú, lo que lo convierte en una herramienta clave para evaluar la generalización de los modelos en contextos globales.

TABLA 2
CANTIDADES DE LOS LABELS MULTILENGUAJE, TAREA 1

Tipo	Entrenamiento	Validación
Humano	257,968	110,166
Generado	416,115	178,728
Total	674,083	288,894

En la siguiente tabla se muestra la distribución de muestras por idiomas, destacándose por su gran mayoría de textos en inglés.

TABLA 3
MUESTRAS POR IDIOMAS

Idioma	Muestras
Inglés	610676
Chino	35284
Búlgaro	8091
Alemán	4693
Italiano	4174
Indonesio	3976
Urdú	3761
Árabe	2114
Ruso	1314

Se presenta un fragmento del dataset multilingüe de la tarea 1.

	id	source	sub_source	lang	model	label	text
0	ec5f3	m4gt	arxiv	en	gemma-7b-it	1	This report summarizes the findings of the US ...
1	5432	mage	wp	en	human	0	I've been standing here for days now. Watching...
2	acd8	mage	xsum	en	flan_t5_xl	1	Towell, 25, was knocked down twice during the ...
3	0d36	hc3	open_qa	zh	gpt-35	1	品牌，它的产品包括化妆品、香水和时装。黑手党是...
4	e32c	hc3	reddit_eli5	en	gpt-35	1	Sometimes when we eat certain types of food, o...

Fig. 3 Muestra del dataset multilenguaje.

El dataset en inglés [18] de la Tarea 2, proviene de ensayos académicos, donde apreciamos una pequeña cantidad de muestras en comparación con el de la tarea 1.

TABLA 4
CANTIDADES DE LA ETIQUETA OBJETIVO EN INGLES, TAREA 2

Tipo	Entrenamiento	Validación
Humano	629	1235
Generado	1,467	391
Total	2,096	1.626

Se incluye una muestra de este dataset.

	id	essay	label
0	168...	I disagree with the statement that the develop...	ai
1	e6a...	I disagree with the statement that the primary...	ai
2	6a0...	International sports events require the most w...	human
3	4d4...	While some individuals may argue that working ...	ai
4	3abf...	I disagree with the statement that working rem...	ai

Fig. 4 Muestra del dataset inglés, tarea 2

El dataset en árabe [19], está compuesto por ensayos académicos escritos por humanos y textos generados automáticamente que provienen de fuentes académicas reconocidas, asegurando autenticidad y calidad en los contenidos. Los textos generados fueron producidos por modelos avanzados, lo que garantiza una representación diversa y alineada con las complejidades lingüísticas del árabe.

TABLA 5
CANTIDADES DE LA ETIQUETA OBJETIVO EN ARABÉ, TAREA 2

Tipo	Entrenamiento	Validación
Humano	1,145	299
Generado	925	182
Total	2,070	481

Seguidamente, se presenta un ejemplo del dataset en árabe.

	id	essay	label
0	ib0eb0d...	الفقه الإسلامي هو مجال دراسي مثير للاهتمام وال...	ai
1	8d92bfb...	في يوم الجمع لم أذه\n:ما فعلت في عطلتي\nيوميات...	human
2	IAS_W_C	ذهبنا الى الحج في البيت الحرام والحج يتطلب تصر...	human
3	IAS_W_C	بسم الله الرحمن الرحيم والصلاة والسلام على نبي...	human
4	i328215...	الهندسة الب\n\nالهندسة البترولية: شغفي الدراسي...	ai

Fig. 5 Muestra del dataset arabé

B. *Preprocesamiento*

El preprocesamiento de datos es una etapa crucial en sistemas de aprendizaje automático, ya que asegura la consistencia, calidad y relevancia de la información al eliminar elementos irrelevantes y preparar los datos para la extracción de características. En este proyecto, se realizó un preprocesamiento adaptado a las particularidades de los datasets, garantizando condiciones óptimas para su análisis.

En los datasets de textos en inglés y árabe, se codificaron las etiquetas "Human" y "AI" en un esquema binario (0 para textos humanos y 1 para generados), facilitando la clasificación eficiente y minimizando errores durante el entrenamiento y evaluación.

Para el dataset multilingüe, se implementó *One-Hot Encoding* para codificar individualmente cada idioma, asegurando que los modelos de aprendizaje automático pudieran reconocer las particularidades lingüísticas de cada uno.

Además, se llevó a cabo una limpieza exhaustiva de los textos en todos los datasets, incluyendo eliminación de elementos irrelevantes como emojis, hashtags, referencias y enlaces, garantizando que los modelos trabajaran exclusivamente con información relevante y de calidad.

Se aplicó lematización para reducir las palabras a su forma base, eliminando variaciones morfológicas y mejorando la consistencia semántica, como por ejemplo corriendo-correr, jumping-jump. Se utilizó spaCy [20] y Stanza [21] para obtener los lemmas según los idiomas de cada dataset.

En la tabla 6 podemos ver los modelos para la obtención de los lemmas en distintos idiomas.

TABLA 6
MODELOS DE LEMATIZACIÓN UTILIZADOS

Lenguaje	Herramienta	Modelo Utilizado
Inglés	spaCy	en_core_web_sm
Alemán	spaCy	de_core_news_sm
Italiano	spaCy	it_core_news_sm
Ruso	spaCy	ru_core_news_sm
Indonesio	Stanza	stanza.Pipeline('id')
Búlgaro	Stanza	stanza.Pipeline('id')
Árabe	Stanza	stanza.Pipeline('ar')
Urdu	Stanza	stanza.Pipeline('ar')
Chino	Stanza	stanza.Pipeline('zh')

En Figura 8 se observa una muestra del dataset lematizado.

	essay	label	texto_limpio
	I disagree with the statement that the develop...	1	I disagree with the statement that the develop...
	I disagree with the statement that the primary...	1	I disagree with the statement that the primary...
	International sports events require the most w...	0	international sport event require the most wel...
	While some individuals may argue that working ...	1	while some individual may argue that work for ...
	I disagree with the statement that working rem...	1	I disagree with the statement that work remote...

Fig. 6 Resultado de la lematización

C. *Extracción de características*

Con los datos preprocesados y lematizados, se procedió a la etapa de extracción de características, una fase crucial en la preparación de los datos para los modelos de aprendizaje automático. Este proceso tiene como objetivo transformar los textos en representaciones numéricas para que los modelos de clasificación binaria puedan interpretarlos. La extracción de características permite resaltar aspectos distintivos del lenguaje humano y generado automáticamente, facilitando que los modelos interpreten y analicen las diferencias subyacentes entre ambos tipos de texto de manera precisa y eficiente. Para esta etapa se extrajeron las siguientes características: N-gramas para TDIF, Perplejidad y Características fraseológicas.

1) *N-gramas para TDIF:*

Para analizar los textos de manera más profunda y efectiva, se utilizó el método TF-IDF combinado con n-gramas (unigrama, bigrama y trigrama), lo que permitió identificar tanto las palabras más relevantes como las combinaciones de palabras que ofrecen un contexto más detallado [22]. El TF-IDF asigna mayor peso a términos que son frecuentes en un texto, pero poco comunes en otros, descartando palabras genéricas, mientras que los n-gramas capturaron relaciones entre palabras consecutivas, proporcionando información sobre patrones contextuales y estilísticos. Este cálculo se realizó para palabras y signos de puntuación.

Se evaluaron configuraciones con límites de 500, 1000 y 1500 tokens. Los resultados indicaron que el límite de 500 tokens ofreció el mejor desempeño. Figura 9 ilustra el cálculo de n-gramas utilizando TF-IDF.

id	!	"	#	$	%	'	(	)	*	...	{	}	-	.	–	’	°	,
6ad	0.0	0.035294	0.0	0.0	0.0	0.039132	0.007621	0.007594	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
38b	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
e6d	0.0	0.000000	0.0	0.0	0.0	0.000000	0.054144	0.053952	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
eb3	0.0	0.403660	0.0	0.0	0.0	0.000000	0.278932	0.277939	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9e8	0.0	0.000000	0.0	0.0	0.0	0.142469	0.277471	0.276482	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig. 7 Resultado de TF-IDF

2) *Características Fraseológicas:*

Se incluyó la extracción de características fraseológicas, esenciales para identificar patrones estilísticos y estructurales en los textos. Estas características se enfocan en aspectos como la longitud promedio de las frases, la complejidad sintáctica, y la frecuencia de ciertas estructuras gramaticales o constructos específicos. Basándose en los principios establecidos por [23], estas métricas analizan cómo las combinaciones de palabras

contribuyen a la fluidez y significado del texto, proporcionando al modelo información clave para una clasificación robusta.

Las características fraseológicas incluidas abarcan una variedad de métricas que evalúan la estructura textual y el uso del lenguaje. A continuación, se presentan las métricas utilizadas y su descripción:

TABLA 7
MÉTRICAS FRASEOLOGICAS UTILIZADAS

Métrica	**Descripción**
Longitud Promedio de las Palabras	Calcula la longitud promedio de las palabras en un texto.
Diversidad Léxica	Mide la proporción de palabras únicas respecto al total.
Longitud Promedio de las Oraciones	Promedio de palabras por oración, un indicador de la complejidad sintáctica.
Desviación Estándar de la Longitud de las Oraciones	Evalúa la variabilidad en la longitud de las oraciones, lo que refleja la fluidez y estructura del texto.
Longitud Promedio de los Párrafos	Calcula el promedio de palabras por párrafo.
Longitud Total del Documento	Mide el número total de caracteres en un texto.
Palabras por Párrafo	Número de palabras promedio por párrafo, excluyendo puntuación.
Palabras por Texto	Contabiliza el número total de palabras.
Oraciones por Texto	Número total de oraciones en un texto, un indicador clave de la segmentación de la información.
Diferencia Promedio en Longitud de las Oraciones	Evalúa la variación promedio en la longitud entre oraciones consecutivas.

Figura 8 muestra ejemplo de las características fraseológicas obtenidas.

id	MeanWordLen	LexicalDiversity	MeanSentenceLen	StdevSentenceLen	MeanParagraphLen	DocumentLen	WordsPerText
bec5f3	6.076923	57.058824	18.500000	3.556684	61.666667	1219	169
)e5432	3.957806	59.071730	15.588235	7.700717	265.000000	1219	237
)4acd8	4.145833	70.833333	26.000000	3.000000	52.000000	253	48
:d0d36	1.486111	66.666667	82.000000	0.000000	82.000000	198	72
72e32c	4.385417	47.916667	18.166667	4.980518	109.000000	537	96

Fig. 8 Resultados de la extracción de características fraseológicas.

3) Perplejidad:

La métrica de perplejidad se empleó para evaluar la fluidez y coherencia de los textos. Esta técnica permitió medir qué tan bien un texto se ajusta a un patrón lingüístico esperado. Textos con perplejidad baja indican una mayor fluidez y predictibilidad, característica de textos humanos, mientras que valores altos señalaron patrones repetitivos o incoherencias propias de textos generados automáticamente. Este análisis resultó esencial para capturar diferencias sutiles entre ambos tipos de textos, proporcionando una métrica robusta para la clasificación [24].

En la Figura 9 se presenta una muestra de esta métrica aplicado a los datasets.

id	Perplexity
ef0f5e94b5168...	5.385571
d078b64fce6a...	5.363896
78ea5ec2d6a0...	17.583391
a72874a224d4...	5.354668
6332a7ea8abf...	5.681499

Fig. 9 Muestra de la perplejidad calculada

D. Entrenamiento.

El entrenamiento del modelo se diseñó para maximizar su capacidad de distinguir entre textos generados automáticamente y escritos por humanos, abordando las particularidades de los datasets y asegurando un aprendizaje equilibrado. Para tratar el desequilibrio en las clases, se aplicaron técnicas como smote y undersampling en datasets más grandes para generar instancias sintéticas, y oversampling en datasets más pequeños para igualar la representación de clases.

Los datos fueron escalados utilizando *MinMax Scaler* para normalizar las características numéricas entre 0 y 1. Los valores faltantes se manejaron reemplazándolos con la mediana de las respectivas columnas para garantizar la consistencia en el conjunto de datos. Para el entrenamiento, se evaluaron varios algoritmos tradicionales de clasificación de textos, incluyendo modelos basados en análisis estadísticos y redes neuronales, mostrados en Tabla 8.

TABLA 8
MODELOS DE CLASIFICACIÓN UTILIZADOS

Modelo	**Motivo de Selección**
Random Forest (RF)	Eficiente en datos complejos y robusto.
Logistic Regression (LR)	Simple y efectivo en problemas binarios.
LinearSVC (SVC)	Escalable y rápido en datasets grandes y con alta dimensionalidad.
Decision Tree (DT)	Fácil de interpretar y maneja datos no lineales.
Multinomial Naive Bayes (NB)	Ideal para datos categóricos y representaciones como TF-IDF.
K-Nearest Neighbors (KNN)	Simple y captura relaciones locales en los datos.
Multi-Layer Perceptron (MLP)	Flexible para captar relaciones complejas entre características.
XGBoost (XGB)	Preciso en datos desbalanceados y resistente al sobreajuste.

Para optimizar los resultados, se implementaron técnicas de ensamblado como *Voting* y *Stacking Classifier*, que combinan las predicciones de múltiples modelos para aprovechar sus fortalezas individuales. Estas estrategias permitieron mejorar la robustez y precisión del sistema, asegurando un desempeño sólido en escenarios multilingües y monolingües.

E. Predicción.

La etapa de predicción se centró en evaluar el rendimiento de los modelos mediante la aplicación a los conjuntos de validación, utilizando como métrica principal el F1-macro, también se calcularon precisión, recall y accuracy, las cuales complementaron el análisis al proporcionar una visión más amplia del rendimiento.

Para facilitar la interpretación de los resultados, se generaron matrices de confusión, que permitieron observar la distribución de las predicciones correctas e incorrectas.

III. RESULTADOS.

Los resultados permiten observar cómo cada modelo responde a las complejidades específicas de cada dataset, reflejando la eficacia del enfoque de clasificación propuesto.

A. Dataset en inglés, tarea 1:

En Tabla 9 se aprecian los resultados obtenidos para este corpus de textos.

TABLA 9
EVALUACIÓN DE MODELOS EN INGLÉS, TAREA 1

Modelo	Precisión	Recall	Accuracy	F1-macro
RF	0.894948	0.909959	0.904729	0.900537
KNN	0.823492	0.824146	0.834546	0.823815
LR	0.736481	0.750128	0.745582	0.738167
SVC	0.737504	0.751065	0.746858	0.739335
DT	0.672613	0.677291	0.642590	0.642290
NB	0.653895	0.663386	0.645413	0.642338
MLP	0.874062	0.882601	0.883942	0.877759
XGB	0.884277	0.898983	0.894338	0.889707
VC	0.892496	0.904036	0.902066	0.897230
SC	0.924019	0.931456	0.931265	**0.927394**

El mejor desempeño fue alcanzado por el modelo ensamblado Stacking Classifier, logrando un F1-macro de 0.927394.

La matriz de confusión para Stacking Classifier se presenta a continuación:

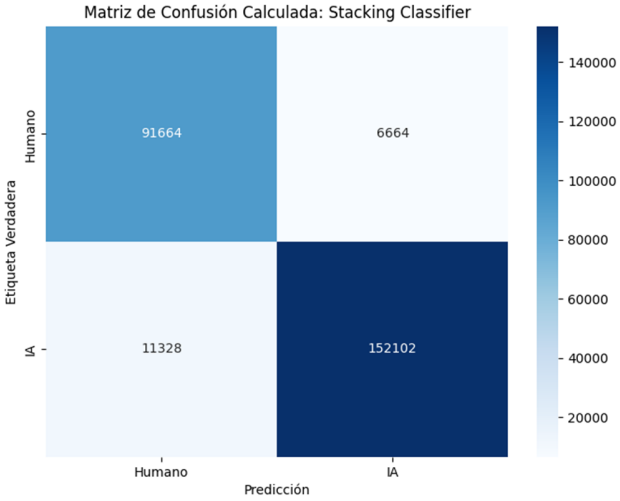

Fig. 10 Matriz de confusión de Stacking Classifier para el dataset en inglés, tarea 1.

B. Dataset multilingüe:

Tabla 10 contiene los resultados para este dataset.

TABLA 10
EVALUACIÓN DE MODELOS EN MULTILENGUAJE, TAREA 1

Modelo	Precisión	Recall	Accuracy	F1-macro
RF	0.904550	0.898638	0.898638	0.898638
LR	0.751127	0.731393	0.725629	0.731393
DT	0.711735	0.748788	0.748123	0.644864
MLP	0.888810	0.885557	0.885557	0.880464
XGB	0.874876	0.888586	0.884224	0.879768
VC	0.889696	0.903101	0.898728	0.894686
SC	0.909149	0.918608	0.917073	**0.913189**

También se observa a Stacking Classifier como el protagonista dentro de este corpus, con un F1-macro de 0.913189.

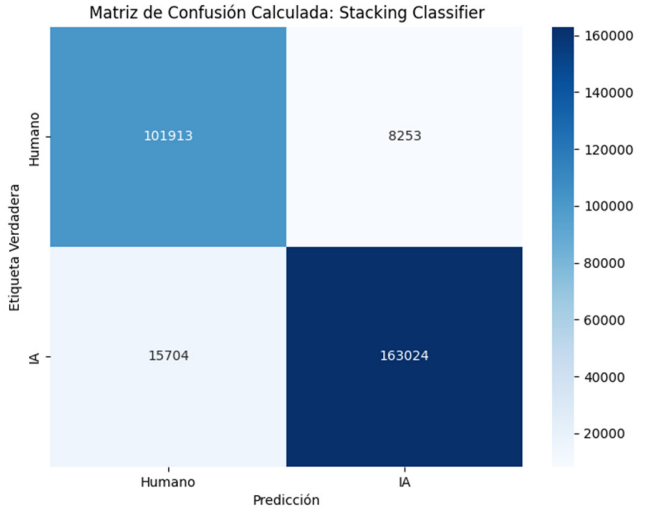

Fig. 11 Matriz de confusión de Stacking Classifier para el dataset en multilingüe.

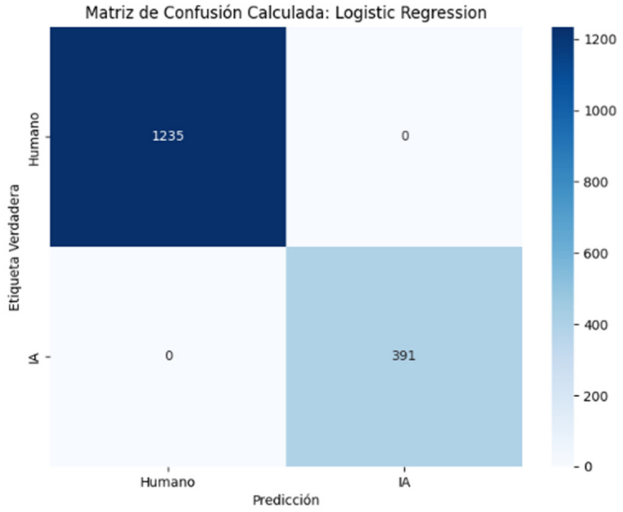

Fig. 12 Matriz de confusión de Logistic Regression para la tarea 2.

C. *Dataset en inglés, tarea 2:*

TABLA 11
EVALUACIÓN DE MODELOS EN INGLÉS, TAREA 2

Modelo	Precisión	Recall	Accuracy	F1-Score
RF	0.979054	0.930946	0.966790	0.952215
LR	1.000000	1.000000	1.000000	**1.000000**
DT	0.951968	0.857716	0.929889	0.893553
MLP	0.998386	0.994885	0.997540	0.996621
XGB	0.960502	0.887127	0.944034	0.917085
VC	0.98695	0.965942	0.982780	0.975917
SC	0.974593	0.929732	0.964076	0.949707

El mejor desempeño fue alcanzado por el modelo Logistic Regression, logrando un F1-macro perfecto del 1.00.

La matriz de confusión para Logistic Regression se presenta a continuación:

D. *Dataset en árabe:*

TABLA 12
EVALUACIÓN DE MODELOS EN ARABÉ, TAREA 2

Modelo	Precisión	Recall	Accuracy	F1-Score
RF	0.997268	0.998328	0.997921	0.997793
LR	0.961929	0.974916	0.968815	0.967347
DT	0.958042	0.969422	0.964657	0.962924
MLP	0.935407	0.954849	0.943867	0.941830
XGB	1.000000	1.000000	1.000000	**1.000000**
VC	0.97644	0.984950	0.981289	0.980296
SC	0.978947	0.986622	0.983368	0.982468

XGBoost alcanzó un desempeño perfecto con un F1-macro de 1.00, clasificando correctamente todos los textos generados y humanos en este conjunto de datos.

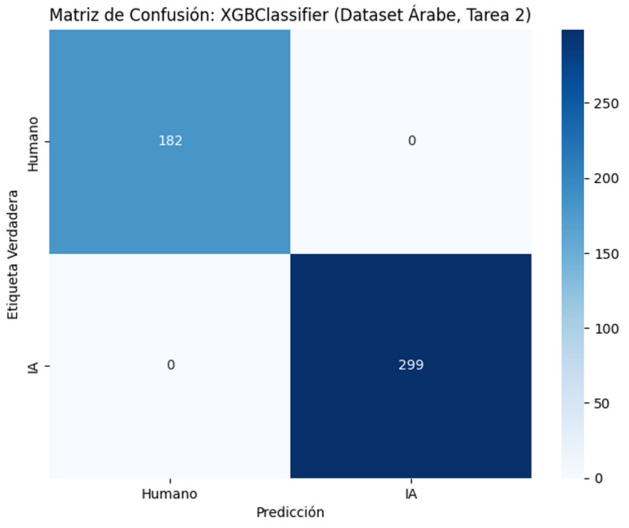

Fig. 13 Matriz de confusión de XGB Classifier

Tabla 13 muestra resumen de los resultados para cada dataset.

TABLA 13
RESUMEN DE LOS MEJORES MODELOS Y RESULTADOS EN LOS DIFERENTES CORPUS.

Dataset	**Modelo**	**F1-macro**
Inglés tarea1	Stacking Classifier	0.927394
Multilingüe	Stacking Classifier	0.913189
Inglés tarea 2	Logistic Regression	1.000000
Árabe	XGBoost	1.000000

IV. DISCUSIÓN

La extracción de características fue fundamental para los resultados obtenidos, evidenciando que cada componente integrado aportó significativamente al rendimiento de los modelos. La combinación de n-gramas, TF-IDF, características fraseológicas y perplejidad permitió analizar los textos desde múltiples niveles, capturando patrones estructurales y de coherencia global esenciales para una clasificación efectiva. Las características fraseológicas destacaron como las más influyentes, justificando su uso prioritario en esta investigación. Además, limitar TF-IDF a 500 tokens demostró ser una decisión acertada, reduciendo el ruido en las muestras y mejorando la precisión al evitar confusiones en los modelos.

En el corpus multilingüe, el modelo Stacking Classifier logró un F1-macro de 0.913, clasificando correctamente 163,024 textos generados y 101,913 textos escritos por humanos, aunque presentó errores al confundir 15,704 textos generados con textos humanos y 8,253 textos humanos con textos generados. Esto podría sugerir que los modelos enfrentan mayores desafíos al identificar textos generados automáticamente en idiomas menos representados, mientras que los textos humanos son predichos con mayor precisión.

De manera similar, en el dataset de gran tamaño en inglés, Stacking Classifier alcanzó un F1-macro de 0.927, clasificando correctamente 152,102 textos generados y 91,664 escritos por humanos. Sin embargo, confundió 6,664 textos humanos como generados y 11,328 generados como humanos. Este rendimiento destaca la capacidad del modelo para manejar grandes volúmenes de datos con alta diversidad estilística y textual, incluso en escenarios complejos.

Por otro lado, los modelos Logistic Regression y XGBoost lograron un desempeño perfecto (F1-macro de 1.0) en datasets pequeños. Logistic Regression clasificó correctamente 1,235 textos humanos y 391 generados en inglés, mientras que XGBoost identificó sin errores 229 textos generados y 182 humanos en árabe. Este nivel de precisión podría atribuirse a la naturaleza homogénea y estructurada de los corpus académicos, que facilita la identificación de patrones distintivos. Sin embargo, esta correlación no puede establecerse con total certeza.

Las diferencias de desempeño entre datasets pequeños frente a corpus grandes y diversos resaltan la necesidad de desarrollar enfoques más avanzados para mejorar la adaptabilidad de los modelos en entornos variados y complejos.

A. *Implicaciones éticas y adaptabilidad del enfoque*

Los hallazgos de esta investigación no solo demuestran la efectividad técnica del enfoque propuesto, sino que también abren la puerta a reflexiones más amplias sobre su aplicabilidad en un panorama tecnológico en constante evolución. En particular, la creciente sofisticación de los modelos generativos plantea desafíos inéditos para la detección de contenido sintético, especialmente cuando estos modelos logran emular con gran precisión los patrones lingüísticos humanos. En este escenario, el uso de características fraseológicas y métricas de coherencia como la perplejidad se presenta como una estrategia robusta para enfrentar estas nuevas complejidades. Sin embargo, el desarrollo y aplicación de estos métodos debe ir acompañado de una reflexión ética profunda. La posibilidad de identificar textos generados o incluso inferir la autoría de textos humanos implica riesgos relacionados con la privacidad, el consentimiento y el uso indebido de los resultados, particularmente en contextos sensibles como el educativo, el legal o el editorial. Por tanto, resulta indispensable que futuros trabajos no solo optimicen la precisión de los modelos, sino que también contribuyan a establecer marcos normativos que aseguren un uso transparente, responsable y respetuoso de estas tecnologías.

V. CONCLUSIÓN

Se desarrolló un método muy efectivo para la identificación de textos generados automáticamente en distintos idiomas, empleando técnicas de extracción de características como n-gramas, TF-IDF, perplejidad y métricas fraseológicas. Estas técnicas permitieron capturar patrones textuales clave tanto en contextos monolingües como multilingües, destacando

la importancia de analizar los textos desde diferentes niveles estructurales y estilísticos.

Los modelos Logistic Regression y XGBoost demostraron un desempeño perfecto, con un F1-macro de 1.0 en datasets pequeños, compuestos principalmente por textos académicos en inglés y árabe. Por su parte, el ensamblado Stacking Classifier mostró un rendimiento sólido en corpus más complejos y diversos, alcanzando un F1-macro de 0.913 en el dataset multilingüe y 0.927 en un corpus en inglés de gran tamaño. Estos resultados evidencian la capacidad de los modelos para manejar datos con alta variabilidad estilística y volumétrica, así como su potencial para identificar patrones textuales distintivos.

Esta investigación destaca la importancia de las características textuales en el rendimiento de los modelos, pero también identifica áreas para futuras mejoras. Es crucial profundizar en el análisis semántico para capturar el significado contextual y evaluar el desempeño en idiomas menos representados. Además, la integración de modelos más avanzados, como redes neuronales profundas y Transformers, podría mejorar la adaptabilidad en contextos multilingües. Finalmente, la ampliación de los corpus con mayor diversidad de géneros y estilos fortalecería la generalización de los modelos.

En resumen, este trabajo establece una base sólida para investigaciones futuras, resaltando la importancia de las características textuales y proponiendo nuevas estrategias para mejorar la detección de textos generados automáticamente. La principal contribución de esta investigación es el desarrollo de un marco metodológico innovador que integra técnicas avanzadas de extracción de características con modelos de clasificación robustos, ofreciendo herramientas clave para enfrentar el desafío de identificar contenido automatizado.

REFERENCIAS

[1] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv Neural Inf Process Syst*, vol. 2020-December, May 2020, Accessed: Nov. 12, 2024. [Online]. Available: https://arxiv.org/abs/2005.14165v4

[2] R. Thoppilan *et al.*, "LaMDA: Language Models for Dialog Applications," Jan. 2022, Accessed: Nov. 12, 2024. [Online]. Available: https://arxiv.org/abs/2201.08239v3

[3] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, Accessed: Nov. 12, 2024. [Online]. Available: https://arxiv.org/abs/2302.13971v1

[4] "Textos generados con IA: La amenaza, la responsabilidad y la promesa." Accessed: Jan. 09, 2025. [Online]. Available: https://latam.turnitin.com/blog/textos-generados-con-ia-amenaza-responsabilidad-promesa

[5] C. Espin-Riofrio, J. L. Charco, D. K. Preciado-Maila, L. Ramos-Ramírez, H. Camacho-Villalva, and A. Montejo-Ráez, "Embeddings of Initial Tokens from BERT-Based Models to Identify Human-Written or Automatically Generated Text," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, Latin American and Caribbean Consortium of Engineering Institutions, 2024. doi: 10.18687/LACCEI2024.1.1.108.

[6] M. D. Bustamante-Rodríguez, A. A. Piedrahita-Ospina, and I. M. Ramírez-Velásquez, "Modelo para detección automática de errores léxico-sintácticos en textos escritos en español," *TecnoLógicas*, vol. 21, no. 42, pp. 199–209, May 2018, doi: 10.22430/22565337.788.

[7] Y. Zhang *et al.*, "Enhancing Text Authenticity: A Novel Hybrid Approach for AI-Generated Text Detection," *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information, ICETCI 2024*, pp. 433–438, Jun. 2024, doi: 10.1109/ICETCI61221.2024.10594194.

[8] S. Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF."

[9] L. Breiman, "Random Forests," 2001.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[11] R. Gu and X. Meng, "AISPACE at SemEval-2024 task 8: A Class-balanced Soft-voting System for Detecting Multi-generator Machine-generated Text," Apr. 2024, Accessed: Nov. 18, 2024. [Online]. Available: https://arxiv.org/abs/2404.00950v1

[12] D. Nguyen, K. M. N. Naing, and A. Joshi, "Stacking the Odds: Transformer-Based Ensemble for AI-Generated Text Detection," Oct. 2023, Accessed: Nov. 18, 2024. [Online]. Available: https://arxiv.org/abs/2310.18906v1

[13] J. J. Sanchez-Medina, "Sentiment analysis and random forest to classify LLM versus human source applied to Scientific Texts," Apr. 2024, Accessed: Nov. 18, 2024. [Online]. Available: https://arxiv.org/abs/2404.08673v1

[14] C. Espin-Riofrio, J. Ortiz-Zambrano, and A. Montejo-Ráez, "An approach to lexicon filtering for author profiling," *Procesamiento del Lenguaje Natural*, no. 71, pp. 75–86, Sep. 2023, doi: 10.26342/2023-71-6.

[15] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-Generated Text be Reliably Detected?," Mar. 2023, Accessed: Dec. 01, 2024. [Online]. Available: https://arxiv.org/abs/2303.11156v3

[16] "Jinyan1/COLING_2025_MGT_en · Datasets at Hugging Face." Accessed: Jan. 09, 2025. [Online]. Available: https://huggingface.co/datasets/Jinyan1/COLING_2025_MGT_en

[17] "Jinyan1/COLING_2025_MGT_multingual · Datasets at Hugging Face." Accessed: Jan. 09, 2025. [Online]. Available: https://huggingface.co/datasets/Jinyan1/COLING_2025_MGT_multingual

[18] "ETS Corpus of Non-Native Written English - Linguistic Data Consortium." Accessed: Jan. 10, 2025. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2014T06

[19] "Arabic Learner Corpus - Linguistic Data Consortium." Accessed: Jan. 10, 2025. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2015S10

[20] I. M. M. Honnibal, "spaCy · Industrial-strength Natural Language Processing in Python." Accessed: Jan. 10, 2025. [Online]. Available: https://spacy.io/

[21] Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz, "Biomedical and clinical English model packages for the Stanza Python NLP library," *Journal of the American Medical Informatics Association*, vol. 28, no. 9, pp. 1892–1899, Sep. 2021, doi: 10.1093/JAMIA/OCAB090.

[22] G. M. Emelyanov, D. V. Mikhailov, and A. P. Kozlov, "The TF-IDF measure and analysis of links between words within N-grams in the formation of knowledge units for open tests," *Pattern Recognition and Image Analysis*, vol. 27, no. 4, pp. 825–831, Oct. 2017, doi: 10.1134/S1054661817040058/METRICS.

[23] M. A. K. Halliday and C. M. I. M. Matthiessen, "Halliday's introduction to functional grammar: Fourth edition," *Halliday's Introduction to Functional Grammar: Fourth Edition*, pp. 1–789, Sep. 2013, doi: 10.4324/9780203431269.

[24] "Perplexity of fixed-length models." Accessed: Dec. 27, 2024. [Online]. Available: https://huggingface.co/docs/transformers/perplexity