VOLUMEN 37
XXX Verano De la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Detección de Textos Generados por Inteligencia Artificial Utilizando Deep Learning

Dario Emilio Hernández Loza[1], Mauro Pantoja Gutiérrez[1], Jonathán de Jesús Estrella Ramírez[1], Juan Carlos Gómez Carranza[1]*

[1]Departamento de Ingeniería Electrónica, División de Ingenierías Campus Irapuato-Salamanca, Universidad de Guanajuato
{de.hernandezloza, m.pantojagutierrez, jdj.estrellaramirez, jc.gomez}@ugto.mx

* Autor de correspondencia

Resumen

Introducción
Identificar la autenticidad y autoría humana de textos escritos se ha convertido en una gran problemática debido al uso extensivo de modelos de inteligencia artificial para la generación de textos, el cual está presente en diversos ámbitos, como la educación, investigación y medios de comunicación.

Métodos
Este artículo presenta un estudio sobre modelos de aprendizaje profundo para la correcta detección de textos generados por inteligencia artificial. Para ello, se usa un conjunto de datos de 2,174 documentos, divididos en textos escritos por humanos y textos generados por inteligencia artificial. Se consideraron para su análisis diferentes modelos populares de transformadores, los cuales fueron probados mediante una validación cruzada estratificada de 5 partes (5-*fold cross-validation*), entrenando los modelos en cada parte (cada fold) utilizando sus versiones preentrenadas y aplicando un ajuste fino (*fine tuning*). Los modelos fueron medidos en su desempeño de clasificación utilizando las métricas de exactitud, precisión, exhaustividad y *F1*, además de los tiempos de entregamiento y prueba.

Resultados
Los resultados demuestran que los modelos lograron un buen desempeño, con una exactitud mínima de 0.919 (BERT, en su versión *bert-base-uncased*), y una exactitud máxima de 0.988 (ALBERT, en su versión *albert-base-v2*).

Conclusiones
En general, los modelos tuvieron una mayor tendencia en identificar mejor los textos generados por inteligencia artificial que los escritos por humanos, en donde se presentaron algunas confusiones. Lo anterior subraya la complejidad de esta tarea, lo que incentiva al desarrollo de herramientas esenciales para preservar la integridad de la información en la era digital.

Palabras clave: procesamiento de lenguaje natural, identificación de textos, aprendizaje profundo, clasificación de textos, transformadores.

1. Introducción

En los últimos años, dentro del área de Procesamiento de Lenguaje Natural (PLN), los avances en el desarrollo de modelos de inteligencia artificial (IA) para la generación de textos han llegado a un nivel en el que, a simple vista, no es posible diferenciar con gran certeza entre textos generados por IA y textos escritos por humanos. Algunos de estos modelos son de fácil acceso (ChatGPT, DeepSeek, Gemini, etc.), permitiendo la creación automática de contenido en diferentes ámbitos como la redacción académica o periodística; ayudando incluso en la resolución de problemas matemáticos y computacionales. Sin embargo, el uso masivo de estas herramientas ha traído algunos desafíos éticos, educativos y sociales. La educación, la investigación y los medios de comunicación son algunos de los sectores que se han visto más afectados, debido a la importancia de la autenticidad y la autoría humana. Por ejemplo, en el ámbito académico, identificar textos generados por IA es fundamental para garantizar el aprendizaje, la integridad de las evaluaciones y prevenir el plagio. Otro gran ejemplo es el ámbito de los medios de comunicación, donde es muy importante identificar

comentarios o noticias generadas por IA, con el fin de evitar la difusión de información falsa, así como la manipulación o la introducción de sesgos no transparentes en la divulgación de contenidos.

Ante estos desafíos, el desarrollo y evaluación constante de modelos capaces de discriminar entre textos creados por humanos y aquellos generados mediante IA es un área de rápido crecimiento en PLN. Similar a otras tareas de PLN, la aplicación de métodos basados en aprendizaje profundo ha generado un gran interés entre los investigadores, con el fin de proponer nuevas herramientas de análisis. En este trabajo, se presenta un estudio sobre modelos de aprendizaje profundo basados en transformadores aplicados en la detección de textos generados por IA. Se considera un grupo de siete modelos de transformadores: ALBERT, BERT, ELECTRA, RoBERTa, XLM-RoBERTa y XLNet. En este estudio se utilizan las versiones preentrenadas de estos modelos y se terminan de entrenar utilizando un ajuste fino (f*ine tuning*). Para su correspondiente análisis y prueba se utiliza un conjunto de datos balanceado de 2,174 documentos, dividido en textos generados por IA y textos escritos por humanos en la misma proporción. Internamente, un preprocesamiento basado en tokenización es aplicado sobre los documentos, y los modelos son configurados para trabajar con la mayor longitud posible de 512 *tokens*. Los hiperparámetros de tasa de aprendizaje y número de épocas de cada modelo son ajustados mediante la estrategia de entrenamiento de un ciclo y detención temprana. Cada modelo es evaluado mediante una validación cruzada estratificada de 5 partes (5-*fold cross-validation*) para asegurar una distribución equitativa de categorías y una evaluación robusta. Durante cada *fold*, el desempeño de cada modelo es evaluado mediante las métricas de exactitud, precisión, exhaustividad y F1, además de los tiempos de entrenamiento y prueba.

Posteriormente, se calcula la mediana del desempeño de cada métrica sobre las 5 partes para realizar la comparación de la efectividad y eficiencia entre modelos con respecto a su capacidad de discriminar entre los dos tipos de textos. De acuerdo con los resultados, el modelo ALBERT demostró el mejor desempeño en cada una de las métricas con un valor en todas ellas de 0.988. Por otro lado, el modelo con los resultados más bajos fue BERT con valores de 0.919 en exactitud, exhaustividad y F1; y 0.929 en precisión. En general, los resultados indican que la arquitectura de trasformadores presenta una buena capacidad de clasificación en este tipo de tarea. Finalmente, la principal contribución de este trabajo es la siguiente: un estudio sobre la validación de modelos de aprendizaje profundo basados en transformadores en la tarea de identificación de textos generados por IA.

El resto del artículo está organizado de la siguiente manera: la Sección 2 presenta un resumen de los trabajos relacionados más relevantes. La Sección 3 describe la metodología aplicada para el uso de los modelos basados en transformadores, así como la descripción del conjunto de datos usado. La Sección 4 presenta los resultados obtenidos con cada uno de los modelos considerados y un análisis sobre su comportamiento. Finalmente, la Sección 5 está dedicada a las conclusiones del estudio realizado e ideas potenciales para trabajos futuros.

2. Trabajo relacionado

El reto *Voight-Kampff Generative AI Authorship Verification 2024* fue presentado formalmente mediante una sesión en el programa del CLEF 2024, teniendo como objetivo la identificación de textos escritos por humanos al ser comparados con textos generados mediante IA. Este reto fue fraccionado en dos tareas individuales bajo un enfoque *builder-breaker*. La primer tarea, *Generative AI Authorship Verification Task* @ PAN, enfocada en la parte *builder*, tenía como objetivo desarrollar sistemas con la capacidad de distinguir entre textos escritos por humanos y textos generados por IA. Por otro lado, la tarea *Voight-Kampff* @ ELOQUENT Lab, enfocada en la parte *breaker*, tenía como objetivo investigar nuevos métodos para la generación de textos con el fin de engañar a los sistemas desarrollados. En este reto se presentaron 43 sistemas, los cuales fueron puestos a prueba con un grupo de 70 conjuntos de datos para conocer su eficiencia y robustez. Los conjuntos de datos presentaban una amplia variabilidad respecto a idiomas, tópicos, textos cortos, entre otros. El desempeño de un sistema era calculado a través de una serie de métricas populares como F1 y el área bajo la curva ROC [1, 2].

Una de las características principales que compartían la mayoría de los sistemas fue la aplicación y análisis de modelos basados en transformadores, específicamente el modelo BERT en su forma base o alguna variante como RoBERTa o DeBERTa. La idea de usar este tipo de modelos era para la obtención de representaciones contextuales. En algunos sistemas, los modelos clasificaban de manera individual los textos para luego comparar las predicciones. En otros casos, bajo un paradigma más clásico en la verificación de autoría, los sistemas analizaban pares de textos de la forma: (texto escrito por humano, texto generado por IA) [1, 2].

Dentro de los 43 sistemas, el sistema con el mejor desempeño fue presentado por Tavan [4], basado en una arquitectura de ensambladores mediante Binoculars como detector y los modelos Mistral y LLaMA2 como analizadores. Los parámetros de estos modelos fueron ajustados mediante el método Low-Rank Adaptation para mejorar la capacidad de análisis. El segundo mejor sistema fue presentado por J. Huang [3], usando BERT en combinación con el método de análisis Tri-Sentencia y una función de pérdida PU.

Por otro lado, algunos sistemas tomaron como base un enfoque más clásico de aprendizaje de máquina, obteniendo también buenas posiciones en el ranking de los trabajos presentados. Dentro de estos y ubicado en la tercer posición, fue el sistema presentado por Lorenz [5]. Este sistema empleó una máquina de vectores de soporte (SVM) con kernel lineal como clasificador, donde los textos eran transformados mediante el método TF-IDF, y solo se consideraban los 1,000 términos más frecuentes. Esta arquitectura demostró una gran robustez en textos del lenguaje alemán sin un previo entrenamiento sobre este. La cuarta posición fue para el sistema desarrollado por L. Guo [6], el cual combinaba técnicas de características multitexto con codificadores. Este sistema usó una versión preentrenada de BERT para la extracción de características, y un modelo BiLSTM (Bidirectional Long Short-Term Memory) como el clasificador.

Otros investigadores propusieron arquitecturas para combinar múltiples características basadas en la tokenización, diversidad léxica, entre otras. Por ejemplo, el sistema desarrollado por A. Valdez-Valenzuela [7], el cual se basa en una estructura de dos caminos; el primero mediante grafos de coocurrencia para GNNs (Graph neural networks), y el segundo mediante el uso de BERT para generar incrustaciones de palabras en conjunto a una extracción de características estilométricas. Otro sistema es el presentado por A. Yadagiri [8], basado en la unión de tres subsistemas que funcionan mediante la comparación entre sí. Cada subsistema estaba compuesto por una etapa de procesamiento con análisis lingüístico y un modelo de BERT. R. Qin [9] presentó también un sistema basado en BERT, donde este aplicaba una corrección ortográfica sobre cada conjunto de datos, y usaba la distancia Levenshtein para el entrenamiento de dos modelos BERT; uno de ellos con R-drop para evitar el sobreajuste. Ambos modelos eran unidos mediante un ensamblador que funcionaba mediante la estrategia de votación. Por otro lado, B. Ostrower [10] presentó un sistema que concatenaba las salidas de un modelo RoBERTa, un grafo de dependencia neuronal y un grafo de coherencia, para alimentar a un clasificador XGBoost.

Este trabajo presenta un estudio del comportamiento de modelos basados en transformadores en la identificación de textos generados por IA, donde los modelos actúan como clasificadores al actualizar los pesos de versiones preentrenadas mediante la técnica de ajuste fino.

3. Metodología

La Figura 3.1 muestra el esquema general de la metodología para el estudio presentado en este artículo. El esquema está compuesto por dos bloques principales: entrenamiento y prueba. La fase de entrenamiento inicia con un conjunto de documentos, el cual es dividido en dos subconjuntos: entrenamiento y validación. Estos subconjuntos son preprocesados para poder construir un modelo basado en transformadores al ajustar los hiperparámetros de este, a través del ajuste fino de pesos de una versión preentrenada. Después de este proceso, en la fase de prueba un conjunto de nuevos documentos es también preprocesado con el fin de evaluar el modelo entrenado en la tarea de identificar textos generados por IA utilizando un conjunto de métricas de desempeño.

Este proceso general se repite utilizando una validación cruzada estratificada de 5 partes (*5-fold cross-validation*), es decir, se repite un total de 5 veces. En cada repetición, los conjuntos de entrenamiento y prueba cambian, construyendo y probando así 5 diferentes versiones de cada modelo de transformadores. Este proceso permitiendo evaluar a cada modelo con mayor robustez. El desempeño final de cada modelo se calcula como la mediana de cada métrica de desempeño sobre las 5 partes.

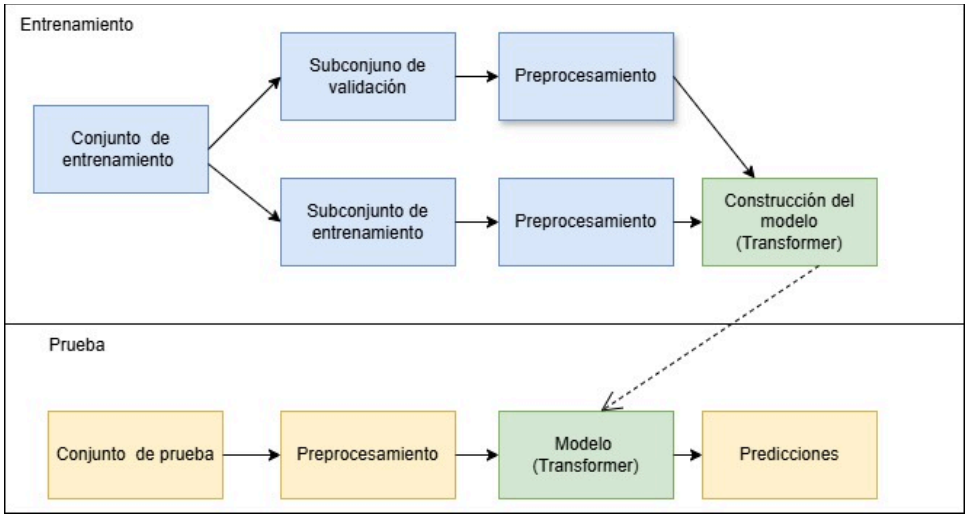

Figura 3.1. Esquema general de la metodología. (Fuente: Elaboración propia)

3.1 Descripción del conjunto de datos

El conjunto de datos original corresponde a la tarea de verificación de autoría generativa en el dominio de noticias, organizada en la conferencia PAN@CLEF 2024[1]. Los documentos dentro de este conjunto están divididos en dos categorías: textos escritos por humanos (1,087 documentos) y textos generados por IA (14,131 documentos); todos ellos escritos en inglés. Para la experimentación en este trabajo, el conjunto de datos fue procesado de la siguiente manera:

1. Extracción del contenido textual de los documentos en el conjunto de datos (organizado originalmente en 13 archivos, conteniendo varios documentos en cada archivo), omitiendo metadatos.
2. Omisión de documentos sin contenido (algunos documentos estaban vacíos de contenido textual).
3. Selección aleatoria de 1,087 documentos generados por IA para formar un conjunto balanceado, obteniendo la misma cantidad de documentos en cada categoría.
4. Los documentos en el conjunto de datos final fueron consolidados en un formato unificado.

La Tabla 3.1 muestra el número de documentos por categoría extraídos para la experimentación, conformando un total de 2,174 documentos. Por otro lado, la Tabla 3.2 muestra un par de documentos de cada una de las categorías, donde se puede observar que los documentos son de longitud variable.

Categoría	Docs.
Humano	1087
IA	1087
Total	2,174

Tabla 3.1. Distribución de documentos por categoría del conjunto de datos usado para la experimentación. (Fuente: Elaboración propia).

[1] Disponible en: https://zenodo.org/records/10718757

Textos generados por humanos	Textos generados por inteligencia artificial
Arizona State University Students Protest Having Kyle Rittenhouse As A Fellow Student Last month's not guilty verdict for Kyle Rittenhouse didn't stop student protesters from gathering on Arizona State University's campus Wednesday to protest…	Russia-Based Hackers Step Up Ransomware Attacks on US Targets, Including Colonial Pipeline Then provide a brief introduction outlining the key points…
Motion seeks evidence of past violence at Kyle Rittenhouse trial KENOSHA, Wis. — Prosecutors in Wisconsin want a judge to allow evidence at Kyle Rittenhouse's trial that shows he had a previous violent encounter in Kenosha before he fatally shot two men and injured another during a police brutality protest last year…	Kamala Harris Takes Precautions with Bluetooth: Is She Paranoid or Just Cautious? In your opinion piece, you may want to take a stance on the issue, for example, you could argue that while Harris may be seen as paranoid by some of her aides, her cautious stance on Bluetooth is warranted…

Tabla 3.2. Ejemplos de textos escritos por humanos (izquierda) y textos generados por IA (derecha). (Fuente: Elaboración propia).

Sobre el conjunto de datos se calcularon una serie de estadísticas con el fin de identificar si los modelos basados en transformadores, que trabajan con una longitud fija de palabras, eran adecuados para abordar este problema. La Tabla 3.3 resume la distribución de palabras por documento en cada una de las categorías, a partir de la media, mediana, desviación estándar, mínimo y máximo de palabras por documento. Las estadísticas revelan que los textos escritos por humanos tienden a estar compuestos por una mayor cantidad de palabras que los textos generados por IA. La desviación estándar sugiere que en ambas categorías el número de palabras es altamente variable entre documentos.

Cantidad de palabras					
Categoría	**Media**	**Mediana**	**Desviación estándar**	**Mínimo**	**Máximo**
Humano	494	475	155	5	1343
IA	409	424	166	10	972

Tabla 3.3. Estadísticas de la cantidad de palabras por documento en cada categoría. (Fuente: Elaboración propia).

Adicionalmente, con el fin de observar la frecuencia del uso de palabras dentro de las categorías, la Figura 3.2 muestra las nubes de palabras obtenidas para cada una; el tamaño de una palabra es proporcional a su frecuencia de uso. En la categoría Humano (textos escritos por personas), las palabras que aparecen con una mayor frecuencia dentro de los textos son "*people*", "*trump*" y "*new*", eso quiere decir que estas aparecen en la mayoría de los documentos. Por otro lado, en la categoría de IA, no hay un conjunto pequeño de palabras que domine a las demás, por lo que en la nube muchas palabras tienen un tamaño similar, por ejemplo "*including*", "*state*", "*continue*", "*social*", "*media*", "*new*" y "*trump*". En general, se puede decir que los humanos tienden a repetir frecuentemente las palabras dentro de un documento, mientras que las herramientas de IA suelen usar palabras con una distribución de frecuencia más uniforme.

Humano IA

Figura 3.2. Nubes de palabras obtenidas de los documentos de cada categoría. (Fuente: Elaboración propia).

3.2 Procesamiento de los datos

Para evaluar la robustez de cada modelo se usa una validación cruzada estratificada de 5 partes (*5-fold cross-validation*), es decir, el conjunto de datos de 2,174 documentos se divide en 5 subconjuntos y se itera 5 veces. Durante cada iteración, se toman 4 subconjuntos para entrenar el modelo y el subconjunto restante se usa para evaluarlo. La técnica para entrenar los modelos basados en transformadores es el ajuste fino, dentro de este, el conjunto de datos de entrenamiento (los 4 subconjuntos) es dividido a su vez en una proporción de 80% para entrenamiento y 20% para validación. Esta subdivisión también se hace de forma estratificada.

El preprocesamiento del conjunto de datos consiste en convertir los documentos en crudo a una representación numérica interpretable por el modelo a evaluar. Este preprocesamiento es aplicado para las tres partes: entrenamiento, validación y prueba. Este método es específico para los modelos basados en transformadores, aplicando las siguientes operaciones:

- **Tokenización:** Los textos se dividen en unidades pequeñas (tokens), como palabras, subpalabras o secuencias de caracteres, utilizando un tokenizador específico del modelo preentrenado.
- **Normalización y estandarización:** Se aplica una serie de reglas para normalizar los tokens obtenidos. En caso de ser necesario, se añaden tokens especiales (como marcadores de inicio de oración o de separación) requeridos por la arquitectura del modelo.
- **Vectorización y alineación dimensional:** Los tokens se convierten en secuencias de identificadores numéricos para después ser concatenadas en un vector de longitud fija. Cuando la cantidad de tokens es menor a esta longitud, se rellena el vector. En el caso de ser mayor, el proceso es truncado, y los tokens restantes son omitidos. La longitud puede ser modificada de acuerdo con cada modelo hasta un límite máximo (aproximadamente 512 tokens).

3.3 Construcción del modelo

El grupo de modelos transformadores considerados para este estudio está compuesto por ALBERT [13], BERT [11], ELECTRA [14], RoBERTa [12], XLM-RoBERTa [16] y XLNet [15]. Se tomaron versiones preentrenadas de cada uno de ellos y posteriormente refinados mediante el ajuste fino. Adicionalmente, se añadió una capa de clasificación en cada uno de los modelos para adaptarlos al tipo de tarea abordada en este trabajo.

Proceso de aprendizaje: El entrenamiento del modelo se llevó a cabo utilizando un enfoque optimizado que ajusta dinámicamente la tasa de aprendizaje a lo largo de las épocas de entrenamiento (*One Cycle Policy*), iniciando con un learning rate de 8.3e-6 y utilizando 7 épocas. Durante este proceso, el modelo aprende a mapear los textos preprocesados a sus etiquetas correspondientes; texto escrito por humano o texto generado por IA.

Criterio de detención temprana: Para prevenir el sobreajuste y optimizar el tiempo de entrenamiento, se implementó un mecanismo de detención temprana. Este mecanismo monitorea el rendimiento del modelo en el subconjunto de validación basado en el área bajo la curva ROC (AUC), deteniendo el entrenamiento si no se observaba mejoras significativas en el número predefinido de épocas, conservando los pesos del mejor ajuste.

Registro de tiempos: Para evaluar la eficiencia de cada modelo, se registró el tiempo total de la fase de entrenamiento.

3.4 Prueba

Después de que un modelo ha sido entrenado, es decir, los pesos de la versión preentrenada han sido ajustados, el paso siguiente es evaluar su capacidad para clasificar nuevos documentos usando el subconjunto de prueba. El conjunto de datos es preprocesado aplicando el mismo método usado con los datos de entrenamiento y validación. Esto asegura que se apliquen idénticas transformaciones (tokenización, normalización, estandarización y alineación dimensional), permitiendo que el modelo reciba datos en el formato exacto que espera y sobre el cual fue entrenado. El modelo entrenado genera predicciones para cada documento del conjunto, dando como salida las etiquetas predichas: 'Humano' para textos escritos por humanos o 'IA' para textos generados por IA.

Medición Detallada del Rendimiento: El desempeño de cada modelo se cuantificó utilizando diversas métricas de clasificación estándar, que proporcionan una evaluación exhaustiva de su precisión y fiabilidad.

Debido a que la tarea abordada es considerada como clasificación binaria ("IA" o "Humano"), se construyó una matriz de confusión para cada parte de la validación cruzada. Esta matriz relaciona las categorías verdaderas de los textos con las categorías predichas por el modelo, ofreciendo una visión detallada de los aciertos y errores. En la Figura 3.3 se observa la representación visual de la matriz de confusión.

		Modelo	
		IA	Humano
Real	IA	**TP**	**FN**
	Humano	**FP**	**TN**

Figura 3.3. Representación de la matriz de confusión. (Fuente: Elaboración propia).

donde TP son los verdaderos positivos (textos de IA clasificados como IA), TN los verdaderos negativos (textos humanos clasificados como humanos), FN los falsos negativos (textos de IA clasificados como humanos) y FP los falsos positivos (textos humanos clasificados como IA). Utilizando la matriz de confusión, se calcularon las siguientes métricas para evaluar el desempeño de cada modelo:

- **Exactitud:** Mide la proporción total de documentos clasificados correctamente entre todos los documentos del conjunto de prueba, véase la Ecuación 3.1.
- **Precisión:** Cuantifica la proporción de documentos clasificados como positivos que realmente pertenecen a esa categoría. Se enfoca en la calidad de las predicciones positivas del modelo, como se muestra en la Ecuación 3.2.
- **Exhaustividad:** Mide cuántos de los documentos positivos fueron correctamente identificados por el modelo. Se centra en la capacidad del modelo para encontrar todas las instancias relevantes. La Ecuación 3.3 define esta métrica.
- **F1-score:** Es la media armónica de la precisión y exhaustividad, proporcionando un equilibrio entre ambas métricas, véase la Ecuación 3.4

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precisión = \frac{TP}{TP + FP} \tag{3.2}$$

$$Exhaustividad = \frac{TP}{TP + FN} \tag{3.3}$$

$$F1 = 2 * \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad} \tag{3.4}$$

Cada una de estas métricas toma valores entre 0 (peor) y 1 (mejor), permitiendo un análisis a profundidad de cada uno de los modelos considerados. Adicionalmente, en cada iteración de la validación cruzada, se registró el tiempo de prueba de

cada modelo. Al evaluar cada modelo mediante validación cruzada, se calculan 5 conjuntos de métricas para cada modelo. Para obtener un resultado final, se calcula la mediana de cada métrica para cada modelo.

4. Resultados

En la Tabla 4.1 se muestran los resultados finales de los modelos de clasificación. Los resultados reportados corresponden a la versión del modelo con el desempeño mediano obtenido en la validación cruzada de 5 partes. En la tabla se muestran las métricas de evaluación, los nombres de las versiones preentrenadas de cada modelo, y los tiempos computacionales (en segundos) para el entrenamiento y la prueba de cada modelo.

De acuerdo con la tabla, el modelo ALBERT alcanzó el mejor desempeño y rendimiento, requiriendo la menor cantidad de tiempo de entrenamiento y logrando valores de 0.988 en todas las métricas. En segundo lugar, fue el modelo RoBERTa con una exactitud de 0.963, una precisión de 0.964, una exhaustividad de 0.963, y un F1 de 0.9631. Ambos modelos demuestran ser bastante fiables, y pueden considerarse una muy buena opción en la tarea abordada. Por otro lado, los modelos ELECTRA y XLNet también mostraron un desempeño aceptable, con valores de 0.951 y 0.942 en F1, respectivamente. Esto indica una buena capacidad para manejar el problema. En contraste, BERT y XLM-RoBERTa se situaron en la parte inferior del espectro de rendimiento, con BERT registrando el F1 más bajo de 0.919, lo que sugiere un margen de mejora en comparación con los otros modelos. Adicionalmente, el modelo XLM-RoBERTa fue el más lento en su entrenamiento.

Modelo	Exactitud	Precisión	Exh.	F1	Entrenamiento (s)	Prueba (s)
ALBERT (base, v2)	**0.988**	**0.988**	**0.988**	**0.988**	**9733**	301
BERT (base, uncased)	0.919	0.929	0.919	0.919	13102	305
ELECTRA (base)	0.951	0.955	0.951	0.951	13994	314
RoBERTa (base)	0.963	0.964	0.963	0.963	11244	**297**
XLM-RoBERTa (base)	0.937	0.940	0.937	0.937	20431	328
XLNet (base)	0.942	0.946	0.942	0.942	19423	5335

Tabla 4.1. Resultados de evaluación de los modelos con validación cruzada estratificada de 5 partes. (Fuente: Elaboración propia).

Como un análisis más detallado del comportamiento de cada modelo, la Figura 4.1 muestra las matrices de confusión obtenidas por cada uno de ellos sobre su respectivo conjunto de prueba. Los valores mostrados en las matrices son la mediana de los resultados de cada parte en la validación cruzada. El modelo ALBERT demuestra un rendimiento notablemente equilibrado. En la detección de textos generado por IA, identificó correctamente 215 instancias frente a 3 clasificadas erróneamente como humanas. De manera similar, para el texto escrito por humanos, acertó en 216 casos, con solo 1 texto atribuido incorrectamente a la IA. Por su parte, el modelo BERT exhibe un perfil asimétrico. Logró una alta precisión al identificar contenido de IA, con 217 aciertos y un único error. Sin embargo, mostró problemas con el texto humano, clasificando correctamente 183 instancias, pero incurriendo en una cantidad considerable de errores al clasificar 34 textos humanos como generados por IA.

Continuando con el análisis, el modelo ELECTRA también revela una tendencia a favorecer la detección de IA. Fue muy eficaz identificando contenido artificial con 216 aciertos y solo 1 error. No obstante, al analizar el texto humano clasificó 197 instancias correctamente, pero atribuyó erróneamente 20 textos humanos a generados por IA. El modelo RoBERTa lleva esta tendencia más allá, mostrando una capacidad casi perfecta para detectar contenido de IA, con 217 ejemplos identificados correctamente y ningún error. En contraste, su discernimiento del texto humano fue más limitado, acertando en 103 instancias, pero clasificando incorrectamente 14 textos humanos como generados por IA. De manera similar, el modelo XML-RoBERTa demostró una gran eficacia con el texto sintético, identificando 217 ejemplos con éxito y un solo error. Sin embargo, también evidenció un punto débil en la clasificación de texto humano, ya que, si bien acertó en 191 casos, atribuyó incorrectamente 26 textos humanos a la IA.

Finalmente, el modelo XLNET presenta un rendimiento excepcional en la detección de contenido artificial, logrando 217 identificaciones correctas sin ningún error. En el análisis de texto humano, acertó en 194 instancias, aunque registró un número significativo de desviaciones, con 23 textos escritos por humanos que fueron erróneamente clasificados como generados por IA.

En general, se puede observar que los modelos tienen una mayor tendencia en identificar textos generados por IA, cometiendo muy pocos errores. En el caso contrario, con la excepción de ALBERT, los modelos cometen una mayor cantidad de errores al clasificar textos escritos por humanos como generados por IA. La Figura 4.3 muestra dos ejemplos

donde los modelos clasifican erróneamente un texto escrito por humano y un texto generado por IA, demostrando la dificultad para categorizarlos.

ALBERT

		Predicha	
		IA	Humano
Real	IA	215	3
	Humano	1	216

BERT

		Predicha	
		IA	Humano
Real	IA	217	1
	Humano	34	183

ELECTRA

		Predicha	
		IA	Humano
Real	IA	216	1
	Humano	20	197

RoBERTa

		Predicha	
		IA	Humano
Real	IA	217	0
	Humano	14	103

XML- RoBERTa

		Predicha	
		IA	Humano
Real	IA	217	1
	Humano	26	191

XLNET

		Predicha	
		IA	Humano
Real	IA	217	0
	Humano	23	194

Figura 4.1. Matrices de confusión de cada modelo entrenado. (Fuente: Elaboración propia).

a. Trump's Social Media Ban: How it Affected the President's Engagement on Facebook and Twitter After the ban, Trump's engagement dropped on both platforms, but some statements have matched or exceeded pre-ban median engagement on other social media accounts and right-leaning platforms…

b. Baldwin's Phone Seized in 'Rust' Shooting Probe, Emails Reveal Gun Choice Discussions Santa Fe, New Mexico - A search warrant has been issued to acquire Alec Baldwin's cell phone for a forensic download, as part of the ongoing investigation into the fatal shooting on the set of the Western film ""Rust."" According to an affidavit from Detective Alexandria Hancock, Baldwin discussed his choice of weapon with Rust armorer Hannah Gutierrez-Reed via email before the incident…

Figura 4.2. Documentos donde los modelos clasifican erróneamente. a) Un texto generado por IA clasificado como texto escrito por humano, y b) Un texto escrito por humano clasificado como texto generado por IA. (Fuente: Elaboración propia).

5. Limitaciones del Estudio

- En el trabajo realizado se limita a un conjunto de 2174 textos, esto debido a la baja cantidad de textos humanos se tuvo que limitar a la poca diversidad de textos generados por Inteligencia artificial.
- El análisis se centró en métricas clásicas (Exactitud, precisión, Exh, tiempos de entrenamiento y prueba y F1), quedando pendiente la exploración de otras métricas que podrían enriquecer la evaluación
- Los resultados obtenidos pueden no generalizarse a textos en otros idiomas, dado que los datos empleado es específico del idioma inglés.

6. Conclusiones

En este trabajo se realizó un estudio para analizar el comportamiento de modelos basados en transformadores sobre la tarea de identificación de textos generados por IA. Para este estudio, se analizó un grupo de seis modelos transformadores con diferentes arquitecturas: ALBERT (base, v2), BERT (base, uncased), ELECTRA (base), RoBERTa (base), XLM-RoBERTa (base) y XLNet (base). El entrenamiento de estos modelos fue mediante el ajuste fino de versiones preentrenadas. Como datos para experimentación, se usó un conjunto procesado de 2,174 documentos perteneciente al conjunto *pan24-generative-authorship-news*, correspondiente a la tarea de verificación de autoría generativa en el dominio de noticias, organizada por PAN@CLEF 2024. Los documentos dentro del conjunto procesado estaban divididos en dos categorías de forma balanceada; textos escritos por humanos y textos generados por IA. Para validar la eficiencia de cada modelo, se aplicó una validación estratificada de 5 partes. El desempeño de cada modelo en cada iteración fue medido mediante las métricas de exactitud, precisión, exhaustividad y F1, además de los tiempos de entrenamiento y prueba.

Los resultados de los experimentos mostraron lo siguiente:

1. ALBERT (base, v2) fue el modelo con el mejor desempeño general, alcanzando un 98.8% en todas las métricas, superando a los demás modelos.
2. Todos los modelos de transformadores mostraron un buen desempeño, con ALBERT destacándose significativamente, seguido por otros modelos como RoBERTa y ELECTRA obteniendo resultados sólidos. Esto indica que las arquitecturas de transformadores son altamente efectivas para diferenciar entre textos creados por humanos de aquellos generados por IA.
3. En cuanto a los errores (documentos mal clasificados), la matriz de confusión de ALBERT mostró que es muy preciso, ya que clasificó incorrectamente solo 3 textos de IA como humanos y solo un texto humano como IA. Otros modelos, como BERT tuvieron más dificultades en clasificar correctamente los textos humanos (34 clasificados incorrectamente como IA). El modelo RoBERTa tuvo 14 textos humanos clasificados incorrectamente como IA.
4. En general, los modelos presentaron una mayor dificultad en detectar los textos escritos por humanos, eso quiere decir, que algunas herramientas de IA han sido capaces de seguir un patrón de escritura muy similar al de humanos.

Hay diferentes caminos interesantes para trabajo a futuro. Primero, se podrían explorar otras arquitecturas de aprendizaje profundo no abordadas en este estudio, incluyendo la exploración de técnicas de ensambladores. Esto también incluye usar conjuntos de datos más grandes y variados para entrenar los modelos, con el fin de mejorar la capacidad de detección. Finalmente, se podría explorar la integración de características contextuales adicionales o datos multimodales (como imágenes o videos que acompañan al texto), las cuales pueden ofrecer nuevas vías para abordar este problema.

Bibliografía/Referencias

[1] PAN at CLEF 2024 - Generative AI authorship verification. (s/f). Webis.de. https://pan.webis.de/clef24/pan24-web/generated-content-analysis.html.

[2] Bevendorff, J., Wiegmann, M., Karlgren, J., Dürlich, L., Gogoulou, E., Talman, A., ... & Stein, B. (2024). Overview of the “voight-kampff” generative AI authorship verification task at PAN and ELOQUENT 2024. In 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble, France 9 September 2024 through 12 September 2024 (Vol. 3740, pp. 2486-2506). CEUR-WS.

[3] Huang, J., Chen, Y., Luo, M., & Li, Y. (2024). Generative AI authorship verification of tri-sentence analysis base on the bert model. Working Notes of CLEF.

[4] Tavan, E., & Najafi, M. (2024). MarSan at PAN: BinocularsLLM, Fusing Binoculars’ Insight with the Proficiency of Large Language Models for Machine-Generated Text Detection.

[5] Lorenz, L., Aygüler, F. Z., Schlatt, F., & Mirzakhmedova, N. (2024). BaselineAvengers at PAN 2024: often-forgotten baselines for LLM-generated text detection. Working Notes of CLEF.

[6] Guo, L., Yang, W., Ma, L., & Ruan, J. (2024). BLGAV: generative AI author verification model based on BERT and BiLSTM. Working Notes of CLEF.

[7] Valdez-Valenzuela, A., & Gómez-Adorno, H. (2024). Team iimasnlp at PAN: leveraging graph neural networks and large language models for generative AI authorship verification. Working Notes of CLEF.

[8] Yadagiri, A., Kalita, D., Ranjan, A., Bostan, A., Toppo, P., & Pakray, P. (2024). Team cnlp-nits-pp at PAN: leveraging BERT for accurate authorship verification: a novel approach to textual attribution. Working Notes of CLEF.

[9] Qin, R., Qi, H., & Yi, Y. (2024). A model fusion approach for generative AI authorship verification. Working Notes of CLEF.

[10] Ostrower, B., Wessell, J., & Bindal, A. (2024). AI authorship verification: an ensembled approach. Working Notes of CLEF.

[11] *Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. arXiv:1810.04805.*

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv:1907.11692.

[13] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*. International Conference on Learning Representations (ICLR). arXiv:1909.11942.

[14] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators*. International Conference on Learning Representations (ICLR). arXiv:2003.10555.

[15] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding*. Advances in Neural Information Processing Systems (NeurIPS). arXiv:1906.08237.

[16] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale (XLM-RoBERTa)*. Proceedings of ACL. arXiv:1911.02116.