Polo del Conocimiento

Pol. Con. (Edición núm. 85) Vol. 9, No 2
Febrero 2024, pp. 998-1010
ISSN: 2550 - 682X
DOI: 10.23857/pc.v9i1

Dataset de textos en español de Ecuador con cuatro versiones reescritas por GPT para tareas de identificación de texto generado automáticamente

Dataset of Spanish texts from Ecuador with four versions rewritten by GPT for automatically generated text identification tasks

Conjunto de dados de textos em espanhol do Equador com quatro versões reescritas pela GPT para tarefas de identificação de texto geradas automaticamente

César Humberto Espín-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

Richard Espinoza-Fajardo [II]
richard.espinozaf@ug.edu.ec
https://orcid.org/0009-0004-5789-1366

Fausto Javier Ortiz-Serrano [III]
fausto.ortizs@ug.edu.ec
https://orcid.org/0009-0001-3965-5253

Tania Peralta-Guaraca [IV]
tania.peraltag@ug.edu.ec
https://orcid.org/0000-0002-4879-6824

Rocio Carchi-Encalada [V]
rocio.carchie@ug.edu.ec
https://orcid.org/0009-0009-6343-2939

Correspondencia: cesar.espinr@ug.edu.ec

Ciencias de la Educación
Artículo de Investigación

*** Recibido:** 30 de diciembre de 2023 ***Aceptado:** 10 de enero de 2024 *** Publicado:** 12 de febrero de 2024

I. Magíster en Sistemas de Información Gerencial, Universidad de Guayaquil, Guayaquil, Ecuador.
II. Universidad de Guayaquil, Guayaquil, Ecuador.
III. Universidad de Guayaquil, Guayaquil, Ecuador.
IV. Magíster en Ingeniería de Software y Sistemas Informáticos, Universidad de Guayaquil, Guayaquil, Ecuador.
V. Máster Universitario en Educación Bilingüe, Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

Los generadores automáticos de texto como GPT de OpenAI, se han vuelto herramientas valiosas por su capacidad de producir texto muy similar al escrito por el humano. Esa capacidad plantea desafíos a la hora de identificar la autoría del texto generado. El enfoque principal del presente trabajo se basa en la necesidad de contar con un dataset de textos en español para ser utilizado en tareas y herramientas de identificación de texto humano o máquina. La intención es proporcionar un dataset de textos en español originario de Ecuador de diversos ámbitos como X (Twitter), noticias y resúmenes de tesis, con una representación variada de estilos y contextos del lenguaje. Utilizando técnicas de web scraping, se recopilaron textos de los distintos dominios, que luego fueron reescritos automáticamente por GPT con la ayuda de la API de OpenAI, generando cuatro versiones distintas de cada uno de los textos originales humanos, para formar así el dataset requerido. De esta manera, se logró formar un conjunto de datos sólido con más de 15,000 textos en español cada uno con su versión original y cuatro versiones reescritas automáticamente por GPT, el mismo que puede ser usado en futuras investigaciones relacionadas a la detección de texto generado automáticamente.

Palabras clave: Generadores automáticos de texto; Dataset, GPT; Procesamiento de Lenguaje Natural.

Abstract

Automatic text generators, such as OpenAI's GPT, have become valuable tools for their ability to produce text very similar to that written by humans. This ability poses challenges when identifying authorship of the generated text. The main focus of this work is based on the need to have a dataset of texts in Spanish to be used in human or machine text identification tasks and tools. The intention is to provide a dataset of texts in Spanish originating in Ecuador from various fields such as X (Twitter), news and thesis summaries, with a varied representation of language styles and contexts. Using web scraping techniques, texts from the different domains were collected, which were then automatically rewritten by GPT with the help of the OpenAI API, generating four different versions of each of the original human texts, thus forming the required dataset. In this way, it was possible to form a solid data set with more than 15,000 texts in Spanish, each with its original version and four versions automatically rewritten by GPT, which can be used in future research related to the detection of automatically generated text.

Keywords: Automatic text generators; Dataset; GPT; Natural Language Processing.

Resumo

Geradores automáticos de texto, como o GPT da OpenAI, tornaram-se ferramentas valiosas pela sua capacidade de produzir texto muito semelhante ao escrito por humanos. Essa habilidade apresenta desafios na identificação da autoria do texto gerado. O foco principal deste trabalho baseia-se na necessidade de contar com um conjunto de dados de textos em espanhol para ser utilizado em tarefas e ferramentas de identificação de textos humanos ou máquinas. A intenção é fornecer um conjunto de textos em espanhol originários do Equador de diversas áreas como X (Twitter), notícias e resumos de teses, com uma representação variada de estilos e contextos linguísticos. Utilizando técnicas de web scraping, foram coletados textos dos diferentes domínios, que foram reescritos automaticamente pelo GPT com o auxílio da API OpenAI, gerando quatro versões diferentes de cada um dos textos humanos originais, formando assim o conjunto de dados necessário. Desta forma, foi possível formar um sólido conjunto de dados com mais de 15.000 textos em espanhol, cada um com sua versão original e quatro versões reescritas automaticamente pelo GPT, que poderá ser utilizado em futuras pesquisas relacionadas à detecção de texto gerado automaticamente.

Palavras-chave: Geradores automáticos de texto; Conjunto de dados; GPT; Processamento de linguagem natural.

Introducción

Actualmente los generadores de texto automático han tenido una evolución significativa, y han experimentado una notoria evolución en su capacidad para producir texto que se asemeja al escrito por humanos. Esto los convierte en herramientas eficientes que pueden ser usados en diversos campos como la educación y el trabajo, ya que pueden generar material de estudio o laboral muy rápidamente y con buenos resultados. Los aspectos principales de la generación de texto según (IEEE Xplore Full-Text PDF: n.d.) es la creación de texto desde cero con la mínima intervención humana y la modificación de texto existente que puede mejorar la claridad, cambiar el tono o adaptar el estilo según requisitos específicos. Entre los generadores de texto más conocidos están: Bard el cual fue lanzado por Google en marzo de 2023 (Manyika, n.d.), Jasper AI un generador de

texto muy versátil y elogiado por los usuarios (Preview & Ai, 2023) y por último tenemos a GPT (Generative Pretrained Transformer) creado por OpenAI, el cual es un modelo de aprendizaje automático que utiliza técnicas de aprendizaje no supervisado y supervisado para comprender y generar lenguaje similar al humano (Lund & Wang, n.d.).

Aunque estas herramientas son útiles, su uso excesivo plantea preocupaciones respecto al impacto en el desarrollo de habilidades fundamentales, como la creatividad y el juicio propio. (Chan, 2023) aborda la creciente inquietud en el ámbito académico respecto al uso de inteligencia artificial generativa de texto y destaca la preocupación por el posible uso indebido de estas herramientas por parte de estudiantes para hacer trampas o copian en sus tareas y exámenes. (Dwivedi et al., 2023) menciona lo útil de ChatGPT en diversos campos. Sin embargo, plantea preocupaciones sobre la dificultad de distinguir entre la autoría humana o máquina. Estas inquietudes sugieren desafíos éticos y prácticos al atribuir la autoría de un texto, afectando la valoración de la originalidad y autenticidad del trabajo humano. (Brown et al., 2020) menciona la creciente dificultad para diferenciar entre texto generado por máquina y humano. Explora beneficios y riesgos, enfocándose en mal uso, sesgos y la capacidad de amplificar actividades perjudiciales.

Por otro lado, cabe mencionar que cada vez es más difícil para el ser humano poder identificar si un texto fue escrito por humano o máquina. (Clark et al., n.d.) aborda la capacidad de personas no expertas para distinguir entre textos generados por inteligencia artificial, especialmente modelos avanzados como GPT, y textos escritos por humanos. (Ippolito et al., n.d.) destaca la limitada capacidad humana para discernir si un texto fue generado automáticamente, incluso expertos en el tema enfrentan dificultades, mostrando una tasa de error del 30% en las evaluaciones realizadas. Todo esto nos indica que hay una necesidad de herramientas capaces de detectar la autoría de texto humano o máquina, esto ha generado que se lleven a cabo investigaciones y tareas que abordan esta necesidad. (Liyanage & Buscaldi, n.d.) nos describe “ALTA” una tarea que involucra la creación de sistemas de detección automática que pueden distinguir entre textos escritos por seres humanos y aquellos generados de forma automática. (Sarvazyan et al., n.d.) nos habla de “AuTexTification”, que forma parte del workshop IberLEF 2023, en donde los participantes, primero tuvieron que reconocer si un texto fue escrito por humano o máquina y después debían atribuir un texto a uno de los seis modelos de generación de textos diferentes. Los resultados generales no fueron concluyentes, pero demostraron que es más fácil detectar el texto en inglés que en español. (LLM - Detect AI Generated Text | Kaggle, n.d.) presentan "LLM - Detect AI

Generated Text", una tarea con el objetivo de simplificar la identificación de textos creados por inteligencia artificial y progresar en el conocimiento actual sobre la detección de modelos de lenguaje de gran tamaño (LLM). (Wu et al., n.d.) proponen large language models detecting (LLMDet), una herramienta de detección de modelos de lenguaje grandes segura y eficiente. No necesita datos de entrenamiento específicos y utiliza información de diversos modelos para identificar textos generados automáticamente. Con una precisión del 98.54% y una velocidad de detección x3.5 más rápida para textos humanos. (Canhasi & Shijaku, n.d.) utilizaron un modelo basado en XGBoost para detectar ensayos generados por ChatGPT, logrando una precisión del 96%. La ingeniería de características fue crucial, destacando la viabilidad de usar aprendizaje automático para identificar texto.

Estas tareas e investigaciones no serían posible sin un conjunto de datos para entrenamiento y prueba, como es el caso de las tareas antes mencionada que utilizaron conjuntos de datos con textos en diferentes idiomas. Existen varios datasets orientados a la detección de texto humano máquina, y mencionamos algunos como: ai-text-detection-pile (Inglés) que contiene 990,000 textos humanos y 340,000 generados por un modelo GPT (Artem9k/Ai-Text-Detection-Pile · Datasets at Hugging Face, n.d.), GPT-wiki-intro (Inglés) contiene 150,00 temas de introducciones de Wikipedia y generados automáticamente por GPT, (Aadityaubhat/GPT-Wiki-Intro · Datasets at Hugging Face, n.d.) y Text sample datasets and AI detectors test results (Inglés) con 100 artículos académicos de acceso abierto sobre salud mental y psiquiatría generados por GPT imitando un estilo académico (Text Sample Datasets and AI Detectors Test Results, n.d.), entre otros.

Es aquí donde se hace presente nuestro trabajo, con un dataset de texto en español de diferentes ámbitos como informal, de noticias y resúmenes de tesis, de origen ecuatoriano, dichos textos serán reescritos por GPT 3.5 el cual nos dará cuatro versiones diferentes del texto original. La recolección de textos fue llevada a cabo mediante técnicas de web scraping, que es, básicamente extraer datos directamente de la web utilizando bots para su posterior análisis (Gomes Barbosa & Cavalcanti, n.d.). los datos se almacenaron en un archivo CSV, para posteriormente ser usado por la API de OpenAi que nos ofrece acceso GPT-3.5, e integrarlo al proyecto. Obtenemos 4 versiones del texto original reescritas por un modelo generador de texto automático, más el texto original. Este dataset final se destinará a futuras herramientas de detección de texto, ya sea humano o generado por máquinas. La diversidad en las reescrituras proporcionará una muestra variada que pueden ser utilizados para entrenar algoritmos que puedan distinguir entre ambos tipos de texto, mejorando

así la capacidad de las aplicaciones futuras para diferenciar el texto escrito por humano del generado automáticamente.

Método

Este trabajo se basa en un análisis bibliográfico documental y cuasi experimental, centrado en la revisión de diversos artículos científicos de gran relevancia y complementado con la exploración de contenido en páginas web especializadas. Este enfoque nos proporcionó una comprensión profunda de diversos métodos para la extracción de datos y generación de texto mediante inteligencia artificial. La recolección de información se automatizó mediante prácticas de web scraping respaldadas por la documentación.

El proceso de desarrollo consta de tres etapas: la evaluación de la estructura de la página para obtener información, la construcción del script para ejecutar el código y la reescritura de todos los textos obtenidos. Recopilamos información de tres dominios diferentes: informal, formal y resúmenes. Para los textos de estilo informal, seleccionamos la red social X (Twitter), mientras que para los textos de estilo formal optamos por un portal de noticias en Ecuador y los resúmenes de tesis fueron recopilados desde el repositorio de la Universidad de Guayaquil. Todos los textos recopilados tienen su origen en Ecuador. El proceso se visualiza en la figura 1.

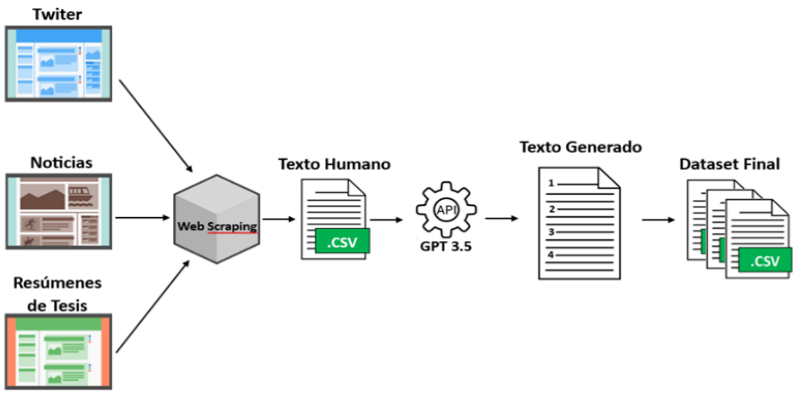


Figura 1: *Método de obtención y rescritura de texto*

Extracción de datos

Implementamos un enfoque integral para la extracción de datos que abarcó tres plataformas clave: X (Twitter), noticias y resúmenes de tesis. Para llevar a cabo este proceso, utilizamos herramientas avanzadas en el ámbito de la programación, destacando el uso de Python y la biblioteca Selenium. En el caso de la recopilación de datos de X, desarrollamos un script utilizando técnicas de web scraping, que interactuó con la interfaz de X. Este script fue diseñado para extraer tweets relevantes, considerando la fecha de creación, el nombre de usuario asociado considerando sea de Ecuador. Esto nos permitió capturar información informal y conversacional, convirtiendo a X en una fuente invaluable de datos que reflejan las opiniones y tendencias actuales en el contexto ecuatoriano. Para el portal de noticias ecuatoriano, optamos por utilizar técnicas de web scraping centradas en la estructura de noticias. Nuestro script, también implementado en Python con Selenium, navegó por secciones específicas del sitio web, extrayendo contenido variado y estructurado. Este enfoque aseguró la obtención de información de calidad, preservando la coherencia y la estructura propia de las noticias, aspectos esenciales para nuestro análisis. En cuanto a los datos académicos, recurrimos al repositorio de tesis de la Universidad de Guayaquil, el script navegó por secciones específicas, capturando los resúmenes de las tesis. Este método nos proporcionó una perspectiva más formal y especializada, complementando la diversidad de datos recopilados en otros contextos.

Generación Automática de texto

En el proceso de reescritura de datos, se empleó la potente API de OpenAI para el modelo GPT-3.5, para transformar los textos originales recopilados de X (Twitter), el portal de noticias y resúmenes de tesis. Además, diseñamos y desarrollamos un script que facilitó la interacción eficiente entre nuestros datos y la API de GPT-3.5. Este script se configuró para gestionar las solicitudes y respuestas de manera efectiva, actuando como un intermediario esencial en el proceso de reescritura de datos. La utilización del script contribuyó a la automatización del proceso de reescritura, agilizando la generación de textos alternativos sin perder la coherencia semántica. Este enfoque integral y automatizado garantizó la eficacia y calidad de las cuatro versiones reescritas, enriqueciendo así nuestro conjunto de datos para análisis posteriores.

Se utilizaron tres parámetros a GPT-3: la instrucción de la solicitud, el mensaje del usuario; el modelo, que será el GPT-3.5-turbo; y la temperatura. La elección del modelo GPT-3.5 se basa en su destacado desempeño en el Procesamiento de Lenguaje Natural, mientras que la temperatura se

Pol. Con. (Edición núm. 85) Vol. 9, No 2, Febrero 2024, pp. 998-1010, ISSN: 2550 - 682X

ajusta para lograr un equilibrio adecuado entre creatividad en la generación y coherencia textual, como se puede observar en la siguiente tabla 1.

Tabla 1: *Parámetros requeridos por el api de OpenAI.*

Parámetro	**Valores**
Modelo	GPT-3.5
Temperatura	0.9

La instrucción de sistema que empleamos es un texto que posibilita su reescritura sin perder el sentido semántico, conservando toda la información, como se evidencia en la figura 2.

```
msj = [
{"role": 'system', "content": """
Eres un asistente que ayuda en la reescritura de textos para dar variedad sin perder el significado  del mensaje, es decir,
para parafrasear textos. Debe conservarse la información relativa a personas, lugares, hashtag , sitios web , referencia , etc.
Proporciona cuatro versiones diferentes enumerándolos y en cada version tiene que estar en una sola linea."""},
{"role": 'user', "content": tweet}
]
```

Figura 2: *Instrucción dada para la generación de texto*

Estos ajustes garantizaron resultados óptimos durante la generación de textos. Es esencial destacar que, antes de enviar los datos a procesar, implementamos un control integral para asegurarnos de que la información recopilada mediante web scraping esté libre de insultos, frases xenófobas o sexistas. Esta medida se tomó para mantener la integridad y ética en la generación de textos, evitando la inclusión de contenido inapropiado. Este enfoque riguroso garantiza que los resultados generados por GPT-3.5 se alineen con los estándares éticos y de calidad que buscamos mantener en nuestra investigación.

Resultados

El resultado obtenido es un dataset con tres conjuntos de datos en español, cada uno correspondiente a diferentes dominios: informal, formal y resúmenes. Los textos fueron recolectados junto con sus respectivas reescrituras, totalizando 15,384 instancias, como se detalla en la siguiente tabla:

Tabla 2: *División de dataset obtenido*

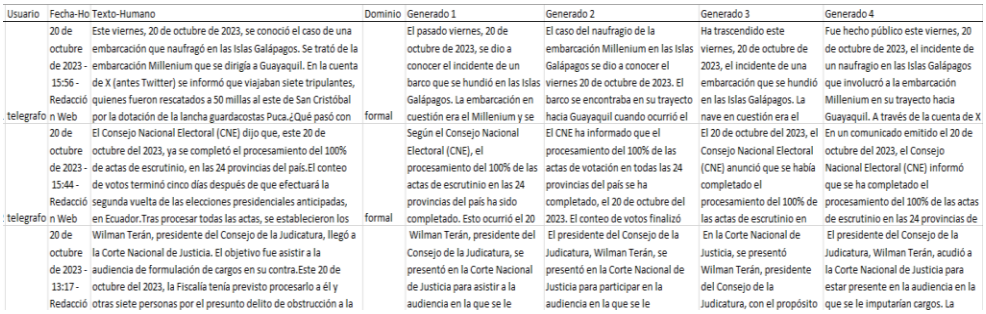
Tipo	**Cantidad**	**Fuente**	**Tiempo extracción (días)**
Informal	5258	X (Twitter)	2
Formal	5053	Noticias	4
Resumen	5073	Resúmenes de Tesis	3
Total	**15384**		**9**

El dataset obtenido, representa un logro significativo en el ámbito del Procesamiento de Lenguaje Natural, proporcionando una base de datos rica y diversa de textos en español. En el dataset de X, se logró la extracción de texto escrito por diversos usuarios ecuatorianos. La diversidad de temas y estilos de escritura en X enriquece el conjunto de datos, permitiendo un análisis más completo del uso del lenguaje en un contexto social y contemporáneo, como se muestra en la siguiente figura.

Usuario	Fecha-Ho	Texto Humano	Dominio	Generado 1	Generado 2	Generado 3	Generado 4
camposemantico_	2023-11-1	Queda muy simple decir: «Te extra	Informal	Es muy básico expresar: "Te echo de menos" cuando Margaret Atwood ya manifestó: "Mi existencia se divide en dos lugares, aquí y donde tú te encuentres".	Resulta demasiado sencillo expresar: "Siento tu ausencia" cuando Margaret Atwood ya expresó: "Mi presencia se desdobla en dos lugares, en este momento y en el sitio donde tú te	Es muy simple decir: "Me haces falta" cuando Margaret Atwood mencionó: "Mi existencia se nutre de dos lugares, de este momento y del lugar donde tú estás".	Resulta bastante básico expresar: "Te extraño" cuando Margaret Atwood ya mencionó: "Mi presencia se desdobla en dos sitios, aquí y en el lugar donde tú te encuentres".
solcasta_	2023-11-0	en tiempos de guerra, todo es válic	Informal	En épocas de conflicto, todas las acciones son aceptables.	Durante períodos de guerra, cualquier medida es legítima.	En tiempos de batalla, todo está permitido.	En situaciones de guerra, todo vale.
				He alcanzado un nivel de autoestima en el que no deseo tener cerca a personas que no me respeten ni me valoren	En este momento de mi vida, tengo suficiente amor propio como para no permitir la presencia de aquellos que no me	Mi nivel de amor propio ha llegado a un punto en el que no quiero tener a personas cerca que no me respeten ni me valoren, a pesar de que	En mi búsqueda del amor propio, he llegado a la conclusión de que no quiero tener cerca a aquellos que no me respeten ni me consideren como yo los considero

Figura 3: *Resultado final de los datos extraídos de X (Twitter) con sus 4 versiones reescritas por GPT.*

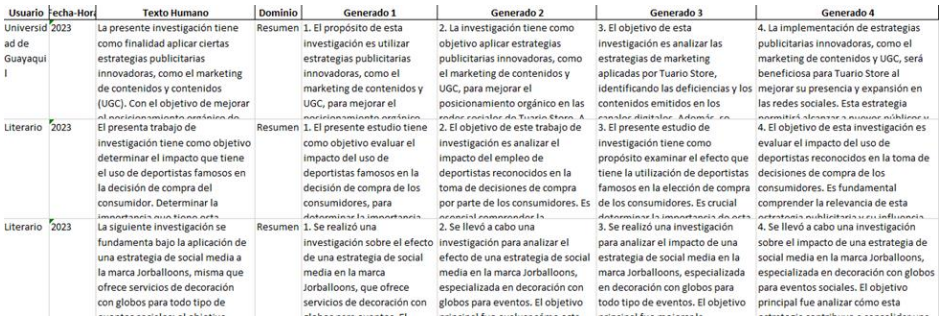
El conjunto de noticias se construyó mediante la obtención de información de un portal de noticias ecuatoriano, refleja el lenguaje utilizado en el ámbito periodístico, capturando la formalidad y la estructura característica de las noticias. La variedad de temas cubiertos proporciona una visión integral del uso del lenguaje en el contexto de la información y las noticias, como se muestra en la siguiente figura.



Usuario	Fecha-Ho	Texto-Humano	Dominio	Generado 1	Generado 2	Generado 3	Generado 4
telegrafo	20 de octubre de 2023 - 15:56 - Redacció n Web	Este viernes, 20 de octubre de 2023, se conoció el caso de una embarcación que naufragó en las Islas Galápagos. Se trató de la embarcación Millenium que se dirigía a Guayaquil. En la cuenta de X (antes Twitter) se informó que viajaban siete tripulantes, quienes fueron rescatados a 50 millas al este de San Cristóbal por la dotación de la lancha guardacostas Puca.¿Qué pasó con	formal	El pasado viernes, 20 de octubre de 2023, se dio a conocer el incidente de un barco que se hundió en las Islas Galápagos. La embarcación en cuestión era el Millenium y se	El caso del naufragio de la embarcación Millenium en las Islas Galápagos se dio a conocer el viernes 20 de octubre de 2023. El barco se encontraba en su trayecto hacia Guayaquil cuando ocurrió el	Ha trascendido este viernes, 20 de octubre de 2023, el incidente de una embarcación que se hundió en las Islas Galápagos. La nave en cuestión era el	Fue hecho público este viernes, 20 de octubre de 2023, el incidente de un naufragio en las Islas Galápagos que involucró a la embarcación Millenium en su trayecto hacia Guayaquil. A través de la cuenta de X
telegrafo	20 de octubre de 2023 - 15:44 - Redacció n Web	El Consejo Nacional Electoral (CNE) dijo que, este 20 de octubre del 2023, ya se completó el procesamiento del 100% de actas de escrutinio, en las 24 provincias del país.El conteo de votos terminó cinco días después de que efectuará la segunda vuelta de las elecciones presidenciales anticipadas, en Ecuador.Tras procesar todas las actas, se establecieron los	formal	Según el Consejo Nacional Electoral (CNE), el procesamiento del 100% de las actas de escrutinio en las 24 provincias del país ha sido completado. Esto ocurrió el 20	El CNE ha informado que el procesamiento del 100% de las actas de votación en todas las 24 provincias del país se ha completado, el 20 de octubre del 2023. El conteo de votos finalizó	El 20 de octubre del 2023, el Consejo Nacional Electoral (CNE) anunció que se había completado el procesamiento del 100% de las actas de escrutinio en	En un comunicado emitido el 20 de octubre del 2023, el Consejo Nacional Electoral (CNE) informó que se ha completado el procesamiento del 100% de las actas de escrutinio en las 24 provincias de
	20 de octubre de 2023 - 13:17 - Redacció	Wilman Terán, presidente del Consejo de la Judicatura, llegó a la Corte Nacional de Justicia. El objetivo fue asistir a la audiencia de formulación de cargos en su contra.Este 20 de octubre del 2023, la Fiscalía tenía previsto procesarlo a él y otras siete personas por el presunto delito de obstrucción a la		Wilman Terán, presidente del Consejo de la Judicatura, se presentó en la Corte Nacional de Justicia para asistir a la audiencia en la que se le	El presidente del Consejo de la Judicatura, Wilman Terán, se presentó en la Corte Nacional de Justicia para participar en la audiencia en la que se le	En la Corte Nacional de Justicia, se presentó Wilman Terán, presidente del Consejo de la Judicatura, con el propósito	El presidente del Consejo de la Judicatura, Wilman Terán, acudió a la Corte Nacional de Justicia para estar presente en la audiencia en la que se le imputarían cargos. La

Figura 4: *Resultado final del dataset de noticias. Elaboración propia.*

En el conjunto de resúmenes de tesis, se recopilaron textos representativos de un estilo más formal y técnico. La variedad de temas abordados en las tesis contribuye a un conjunto de datos que abarca diversas áreas del conocimiento. Este dominio proporciona una perspectiva única sobre el lenguaje utilizado en contextos académicos y científicos en español, como se muestra en la siguiente figura.



Usuario	Fecha-Hora	Texto Humano	Dominio	Generado 1	Generado 2	Generado 3	Generado 4
Universidad de Guayaquil	2023	La presente investigación tiene como finalidad aplicar ciertas estrategias publicitarias innovadoras, como el marketing de contenidos y contenidos (UGC). Con el objetivo de mejorar el posicionamiento orgánico de	Resumen	1. El propósito de esta investigación es utilizar estrategias publicitarias innovadoras, como el marketing de contenidos y UGC, para mejorar el posicionamiento orgánico de	2. La investigación tiene como objetivo aplicar estrategias publicitarias innovadoras, como el marketing de contenidos y UGC, para mejorar el posicionamiento orgánico en las redes sociales de Tuario Store. A	3. El objetivo de esta investigación es analizar las estrategias de marketing aplicadas por Tuario Store, identificando las deficiencias y los contenidos emitidos en los canales digitales. Además, se	4. La implementación de estrategias publicitarias innovadoras, como el marketing de contenidos y UGC, será beneficiosa para Tuario Store al mejorar su presencia y expansión en las redes sociales. Esta estrategia permitirá alcanzar a nuevos públicos y
Literario	2023	El presenta trabajo de investigación tiene como objetivo determinar el impacto que tiene el uso de deportistas famosos en la decisión de compra del consumidor. Determinar la importancia que tiene esta	Resumen	1. El presente estudio tiene como objetivo evaluar el impacto del uso de deportistas famosos en la decisión de compra de los consumidores, para determinar la importancia	2. El objetivo de este trabajo de investigación es analizar el impacto del empleo de deportistas reconocidos en la toma de decisiones de compra por parte de los consumidores. Es esencial comprender la	3. El presente estudio de investigación tiene como propósito examinar el efecto que tiene la utilización de deportistas famosos en la elección de compra de los consumidores. Es crucial determinar la importancia de esta	4. El objetivo de esta investigación es evaluar el impacto del uso de deportistas reconocidos en la toma de decisiones de compra de los consumidores. Es fundamental comprender la relevancia de esta estrategia publicitaria y su influencia
Literario	2023	La siguiente investigación se fundamenta bajo la aplicación de una estrategia de social media a la marca Jorballoons, misma que ofrece servicios de decoración con globos para todo tipo de	Resumen	1. Se realizó una investigación sobre el efecto de una estrategia de social media en la marca Jorballoons, que ofrece servicios de decoración con	2. Se llevó a cabo una investigación para analizar el efecto de una estrategia de social media en la marca Jorballoons, especializada en decoración con globos para eventos. El objetivo	3. Se realizó una investigación para analizar el impacto de una estrategia de social media en la marca Jorballoons, especializada en decoración con globos para todo tipo de eventos. El objetivo	4. Se llevó a cabo una investigación sobre el impacto de una estrategia de social media en la marca Jorballoons, especializada en decoración con globos para eventos sociales. El objetivo principal fue analizar cómo esta

Figura 5: *Resultado final del dataset de resúmenes de tesis. Elaboración propia.*

La aplicación de la API de GPT-3.5 ha demostrado ser altamente efectiva en la generación de múltiples versiones reescritas para cada fragmento de texto, proporcionando una diversidad que se ajusta a distintos tonos y estilos de escritura. La capacidad de preservar la información esencial ha sido clave, garantizando la coherencia y relevancia en diferentes contextos, desde textos informales de redes sociales hasta documentos académicos formales. La herramienta de interacción desarrollada, mediante un script, ha facilitado de manera eficiente la comunicación entre los datos obtenidos y la API, permitiendo un control efectivo sobre el proceso de reescritura. La validación manual ha confirmado la fidelidad al contenido original y la adaptabilidad al contexto, asegurando resultados de calidad. La aplicación de webscraping para la recopilación de datos ha arrojado un conjunto robusto y diversificado, abarcando diferentes tipologías de textos en español, tanto formales como informales. El enfoque integral, que combina la potencia de la inteligencia artificial

con técnicas de obtención de datos, sienta las bases para futuras investigaciones y aplicaciones en el campo del procesamiento de lenguaje natural en español.

Discusión

El dataset obtenido de texto en español de diferentes dominios como X (Twitter), noticias y resúmenes de tesis, fue posible gracias a los métodos utilizados. Es importante mencionar que esto es un enfoque, novedoso no solo porque sea en español y que este compuesto de textos originarios de Ecuador, sino también porque se generaron cuatro versiones del texto original que da variedad al estilo de escritura de GPT, y así poder identificar en trabajos futuros, cómo el modelo estructura y crea una oración.

Si bien se utilizó GPT como modelo de generación de texto automático para la reescritura, se pudieron elegir otros modelos, pero GPT se adaptó a las necesidades del proyecto. Explorar otros modelos y comparar la estructura en la escritura es algo que podría ser posible en futuros trabajos. Si bien el dataset final está completo es su totalidad, se podría aumentar la cantidad de dominios, y poder obtener más variedad y cantidad en los textos.

Este dataset será una gran herramienta que podrá ser utilizado en futuras tareas, investigaciones y herramientas orientadas a la detección de texto automático, por tanto, contribuye a apreciar la originalidad de los textos humanos.

Conclusiones

En esta investigación se usaron métodos y técnicas indispensables para la creación de un dataset de textos en español. La recolección de texto se llevó a cabo mediante técnicas avanzadas de web scraping, permitiendo la extracción de más de 15,000 textos en español procedentes de diversos ámbitos como X (Twitter), noticias y resúmenes de tesis, todos originarios de Ecuador. Estos textos representan una amplia gama de formalidades lingüísticas utilizadas por los usuarios ecuatorianos, ofreciendo así una muestra representativa y diversa del lenguaje en contexto.

Además, se destaca también, la correcta elección del modelo GPT-3.5 y la implementación de la API de OpenAI, que facilitaron la generación coherente y variada de cuatro versiones reescritas de cada texto original obtenido. Esta elección permitió no solo mantener la coherencia en los textos

generados, sino también preservar el sentido original, su diversidad y relevancia en distintos contextos.

Este proceso ha culminado en la creación exitosa de un conjunto de datos no solo valioso en sí mismo, sino que también se posiciona como una herramienta para investigaciones futuras y avances en el campo de la detección de texto generado automáticamente.

Referencias

1. aadityaubhat/GPT-wiki-intro · Datasets at Hugging Face. (n.d.). Retrieved January 22, 2024, from https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro
2. Alves, A. D. (n.d.). Introdução à API da OpenAI. https://platform.openai.com/docs/supported-countries
3. artem9k/ai-text-detection-pile · Datasets at Hugging Face. (n.d.). Retrieved January 22, 2024, from https://huggingface.co/datasets/artem9k/ai-text-detection-pile
4. Canhasi, E., & Shijaku, R. (n.d.). ChatGPT Generated Text Detection. https://doi.org/10.13140/RG.2.2.21317.52960
5. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20(1). https://doi.org/10.1186/s41239-023-00408-3
6. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (n.d.). Human Evaluation of Generated Text. 7282–7296. Retrieved January 8, 2024, from www.nltk.org/
7. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/J.IJINFOMGT.2023.102642
8. Gomes Barbosa, A. B., & Cavalcanti, A. B. (n.d.). Web Scraping e Análise de dados.
9. IEEE Xplore Full-Text PDF: (n.d.). Retrieved December 4, 2023, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10177704

10. Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (n.d.). Automatic Detection of Generated Text is Easiest when Humans Are Fooled. Association for Computational Linguistics. https://github.com/openai/
11. Liyanage, V., & Buscaldi, D. (n.d.). An Ensemble Method Based on the Combination of Transformers with Convolutional Neural Networks to Detect Artificially Generated Text. Retrieved November 24, 2023, from https://gptzero.me/
12. LLM - Detect AI Generated Text | Kaggle. (n.d.). Retrieved January 21, 2024, from https://www.kaggle.com/competitions/llm-detect-ai-generated-text/overview
13. Lund, B. D., & Wang, T. (n.d.). Chatting about ChatGPT: how may AI and GPT impact academia and libraries? https://doi.org/10.1108/LHTN-01-2023-0009
14. Manyika, J. (n.d.). An overview of Bard: an early experiment with generative AI.
15. Preview, A., & Ai, J. (2023). Competitor Analysis Report. https://zapier.com/blog/jasper-ai/
16. Sarvazyan, A. M., José´, J., González, J., Franco-Salvador, M., Rangel, F., Chulvi, B., & Rosso, P. (n.d.). Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains. Retrieved November 24, 2023, from https://tinyurl.com/bloom-1b7
17. Text sample datasets and AI detectors test results. (n.d.). Retrieved January 21, 2024, from https://figshare.com/articles/dataset/Text_sample_datasets_and_AI_detectors_test_results/24208443
18. Wu, K., Pang, L., Shen, H., Cheng, X., & Chua, T.-S. (n.d.). LLMDet: A Third Party Large Language Models Generated Text Detection Tool. https://github.com/TrustedLLM/LLMDet.