OPTIMIZACIÓN DE *DATASETS* DE LA *DARK WEB* PARA SU PROCESAMIENTO EN *OPENAI*

Víctor-Pablo Prado-Sánchez, Adrián Domínguez-Díaz, Luis de-Marcos
y José-Javier Martínez-Herráiz
Departamento de Ciencias de la Computación, Universidad de Alcalá
Alcalá de Henares, España

RESUMEN

Este estudio aborda la optimización del uso de tokens y la reducción de costos operativos en la clasificación de contenidos de la Dark Web utilizando modelos LLM de OpenAI. La metodología aplicada incluyó la limpieza y filtración de datos, lo que condujo a una reducción del 85% en el consumo de tokens en direcciones con documentos en inglés, sin comprometer el rendimiento del modelo GPT-3.5 Turbo en cuanto a precisión, sensibilidad y F1-score. Se concluye que la optimización de tokens es efectiva para reducir recursos y costes sin comprometer la clasificación de grandes volúmenes de datos.

PALABRAS CLAVE

Dark Web, Darknet, Datasets, LLM, OpenAI, ChatGPT

1. INTRODUCCIÓN

El incremento en el volumen y la complejidad de las actividades en la Dark Web ha generado la necesidad de herramientas avanzadas para la clasificación y análisis de sus contenidos, generalmente relacionados con actividades ilegales o dañinas. Para abordar este desafío, se han utilizado modelos de lenguaje preentrenados como los LLMs (Large Language Models) de OpenAI, que presentan un enfoque prometedor para la clasificación automática de textos (Avarikioti et al., 2018). No obstante, el uso de estos modelos en la clasificación de textos de la Dark Web plantea desafíos relacionados con el coste de procesamiento, principalmente debido al uso de tokens (Clavié et al., 2023).

Diversos estudios han abordado la clasificación de contenidos en la Dark Web mediante enfoques supervisados que requieren el etiquetado manual de datos, como es el caso de los conjuntos DUTA (Darknet Usage Text Addresses) (Al Nabki et al., 2017), DUTA-10K, una mejora de DUTA (Al-Nabki et al., 2019) y CoDA (Comprehensive Darkweb Annotations) (Jin et al., 2022). Aunque estos enfoques han demostrado su eficacia, el rendimiento de los modelos supervisados depende de la disponibilidad de datos etiquetados y del entrenamiento específico en cada tarea, lo que dificulta su adaptación a la naturaleza dinámica de la Dark Web (Al Nabki et al., 2017).

Los LLMs de OpenAI presentan una alternativa basada en *zero-shot*, que permite la clasificación sin necesidad de un entrenamiento previo o con un mínimo de ejemplos (Kalyan, 2024). En este contexto, se ha evaluado el rendimiento de GPT-3.5 en la clasificación *zero-shot* de contenidos de la Dark Web, alcanzando un valor F1 ponderado del 80,5%. Sin embargo, se detectaron grandes diferencias entre categorías y ciertas limitaciones que afectan su rendimiento en comparación con clasificadores supervisados (Prado Sánchez et al., 2024). Aunque ChatGPT también ha sido evaluado en estas tareas, los resultados muestran que su fiabilidad es comparable a la de los clasificadores humanos, especialmente en categorías tecnológicas. Sin embargo, las explicaciones generadas por el modelo no siempre mejoran de manera consistente la comprensión humana (Prado-Sánchez et al., n.d.).

La tokenización, proceso de dividir el texto en tokens, impacta tanto en el rendimiento como en los costos de los LLMs. OpenAI cobra en función de los tokens utilizados, lo que hace que el análisis de textos largos, como los de la Dark Web, pueda ser costoso dependiendo del idioma y tipo de texto. Palabras largas generan más tokens, mientras que las cortas generan menos (“OpenAI Platform,” n.d.).

Tabla 1. Relación aproximada entre tokens, palabras y caracteres en textos útil para estimar el uso de tokens en los modelos de lenguaje más recientes disponibles dentro de la API de OpenAI

Tokens	Aproximación en palabras	Aproximación en caracteres (con espacios)
1 tokens	0.75 palabras	4 caracteres
100 tokens	75 palabras	400 caracteres
1.000 tokens	750 palabras	4.000 caracteres

La Tabla 1 ilustra la relación aproximada entre tokens, palabras y caracteres. Esta relación es útil para dimensionar los costos y planificar el uso eficiente de recursos en tareas específicas.

Este análisis de tokenización es relevante en el caso de modelos como GPT-3.5 y GPT-4, se emplean diferentes esquemas de tokenización, que pueden impactar en el rendimiento, particularmente en tareas aritméticas y de procesamiento numérico (Singh and Strouse, 2024). ChatGPT, por ejemplo, utiliza un tokenizador preentrenado que optimiza la segmentación de palabras y subpalabras, mejorando la capacidad del modelo para manejar palabras raras o fuera del vocabulario (Roumeliotis and Tselikas, 2023).

Asimismo, se ha observado que la variabilidad en el número de tokens necesarios para representar información en distintos idiomas genera desigualdades en los costes, un aspecto relevante en el diseño de políticas de precios más equitativas para el uso de estos modelos (Ahia et al., 2023).

Tabla 2. Comparativa de los modelos de lenguaje más recientes disponibles dentro de la API de OpenAI, con su respectiva descripción, capacidad de tokens y costes

Modelo	Descripción	Tokens entrada	Coste entrada (€/token)	Tokens salida	Coste salida (€/token)
GPT-4o	Modelo avanzado optimizado para tareas complejas con gran capacidad de contexto.	128.000 tokens	0.030 €/token	16.384 tokens	0.060 €/token
GPT-4o-mini	Versión más ligera del GPT-4o, adecuado para tareas de menor escala con eficiencia.	128.000 tokens	0.025 €/token	16.384 tokens	0.050 €/token
GPT-3.5-turbo-0125	Modelo optimizado y económico para aplicaciones más ligeras con menor capacidad de contexto.	16.385 tokens	0.015 €/token	4.096 tokens	0.030 €/token

La Tabla 2 presenta una comparativa de los modelos más recientes disponibles en la API de OpenAI, destacando sus capacidades de procesamiento de tokens, costes por token de entrada y salida, así como los datos de entrenamiento utilizados. Se observa que modelos como GPT-4o y GPT-4o-mini manejan hasta 128.000 tokens de entrada, con costes de 0.030 €/token y 0.025 €/token, respectivamente. Modelos como GPT-3.5-turbo-0125 están optimizados para tareas más ligeras, con 16.385 tokens de entrada y un coste de 0.015 €/token. La capacidad de procesamiento y el coste de los tokens afectan directamente la eficiencia y viabilidad económica de utilizar LLMs para el análisis de grandes conjuntos de datos, siendo este análisis económico fundamental para evaluar la sostenibilidad de aplicar estos modelos en entornos reales.

Tras revisar el estado del arte, se observa la ausencia de estudios que analicen el procesamiento de datasets de la Dark Web para su clasificación utilizando los modelos LLM de la API de OpenAI. Este estudio tiene como objetivo analizar los costes asociados al uso de los modelos LLM de OpenAI en la clasificación de contenidos de la Dark Web, así como explorar estrategias para optimizar el uso de tokens. Los objetivos de la investigación son los siguientes:

- OI1: Explorar estrategias para optimizar el uso de tokens en los modelos LLM de OpenAI, con el fin de mejorar la eficiencia del procesamiento y reducir los costes operativos en la clasificación de contenidos.
- OI2: Evaluar el uso de modelos LLM de OpenAI en la clasificación de contenidos de la Dark Web, analizando su impacto en la eficiencia y efectividad del proceso de clasificación tras el procesamiento del conjunto de datos.

Este trabajo aporta conocimiento sobre la relación entre la optimización asociada al uso de los modelos LLM de OpenAI y su rendimiento, identificando estrategias para optimizar el uso de tokens y mejorar la eficiencia del procesamiento.

El resto de este documento se estructura de la siguiente manera. En la Sección 2 se presenta una visión general de los trabajos relacionados con la tokenización de los LLMs, los costos asociados a OpenAI y la justificación para el procesamiento del dataset DUTA-10K. La Sección 3 se detalla la metodología empleada para llevar a cabo procesamiento de datasets de la Dark Web. La Sección 4 presenta los resultados y su discusión, en base al procesamiento realizado sobre el dataset. Finalmente, en la Sección 5 se exponen las conclusiones sobre la investigación llevada a cabo.

2. ESTADO DEL ARTE

2.1 La Tokenización en los *LLMs*

La tokenización se define como el proceso de dividir un texto en unidades más pequeñas llamadas tokens, esenciales para que los modelos de lenguaje procesen información y calculen costes. Este proceso influye en el rendimiento de los modelos y puede generar sesgos inductivos, tanto útiles como perjudiciales (Roumeliotis and Tselikas, 2023).

Históricamente, los LLMs han utilizado codificación por pares de bytes, sin considerar dominios específicos. Con el aumento del uso de LLMs para tareas de razonamiento, se han implementado esquemas de tokenización específicos para números. Por ejemplo, modelos como LLaMa y PaLM tokenizan cada dígito de manera individual, mientras que GPT-3.5 y GPT-4 utilizan tokens (Singh and Strouse, 2024).

La tokenización también es un proceso fundamental en ChatGPT. Este modelo emplea un tokenizador preentrenado que convierte el texto de entrada en una secuencia de tokens, lo que permite al modelo procesar el lenguaje de manera más efectiva. Además, se aplica la codificación de subpalabras para manejar palabras raras o fuera de vocabulario (Roumeliotis and Tselikas, 2023).

El preprocesamiento se considera esencial para determinar la calidad de los datos de entrada, y se destaca la limpieza de datos como un paso crítico para eliminar información irrelevante que puede afectar el rendimiento del modelo (Manning et al., 2014). En el contexto de diferentes idiomas, algunas lenguas pueden requerir hasta cinco veces más tokens que otras para transmitir la misma información (Ahia et al., 2023). Por lo tanto, el diseño del modelo y la tokenización impactan el rendimiento y la equidad en el uso de los modelos de lenguaje.

2.2 Costes Dentro de *OpenAI*

El uso de modelos como ChatGPT impacta en la productividad y los costos operativos, debido al procesamiento basado en tokens. Sin embargo, es importante considerar los costes asociados. Los proveedores de API, como OpenAI, cobran a los usuarios en función del número total de tokens procesados o generados, lo que puede generar disparidades en los costes según el idioma utilizado (Ahia et al., 2023).

Específicamente, ChatGPT tiene un límite máximo de longitud de secuencia de 4096 tokens, que incluye tanto los tokens de entrada como los generados. Esto significa que el número de tokens utilizados puede afectar directamente el coste de uso de la API, ya que los usuarios son cobrados en función del total de tokens procesados (Ahia et al., 2023).

La fragmentación de tokens en ciertos idiomas puede resultar en un uso menos eficiente de la API, incrementando los costes para hablantes de esos idiomas. Además, se subraya la importancia de abordar preocupaciones sobre sesgos y plagio en el uso de ChatGPT, sugiriendo la necesidad de establecer directrices éticas para su aplicación en diversos contextos, incluida la ciencia de datos (Roumeliotis and Tselikas, 2023). La relación entre tokens y palabras se considera crucial, con una media aproximada de 1 token equivalente a 0.75 palabras, lo que puede variar según el tipo de texto y el idioma (“OpenAI Platform,” n.d.).

2.3 Proceso de Limpieza del *Dataset* DUTA-10K

Los estudios sobre el dataset DUTA-10K han explorado técnicas de limpieza y preparación de archivos .txt para su análisis. Estas metodologías, como la eliminación de etiquetas HTML y duplicados, han buscado garantizar datos más consistentes y útiles para la clasificación mediante modelos de lenguaje. El contenido fue extraído mediante el navegador Lynx, que eliminó las etiquetas HTML y retuvo solo el texto visible. Sin embargo, este método dejó elementos redundantes, como enlaces duplicados, que podían afectar el análisis. Para evitar la formación de multigrafos y preservar la integridad del contenido, se procedió a la eliminación de estos duplicados (Al-Nabki et al., 2019).

El sistema de clasificación automática utilizado en DUTA-10K, basado en algoritmos de aprendizaje supervisado y validado mediante etiquetado manual, presentó dificultades como la ambigüedad del lenguaje en los textos, el acceso restringido a ciertos dominios y la variabilidad del contenido en la darknet de Tor. Además, el proceso de etiquetado manual introdujo sesgos que afectaron la consistencia de las etiquetas.

Debido a estos problemas, se determinó que los archivos .txt originales no eran adecuados para un uso directo sin un proceso de limpieza exhaustivo. Este procedimiento permitió eliminar errores y redundancias que podrían distorsionar los resultados del análisis, garantizando un conjunto de datos más depurado y adecuado para su uso en la clasificación de contenidos mediante modelos LLM de OpenAI.

3. METODOLOGÍA

3.1 Procesamiento de *Datasets* de la *Dark Web*

El procesamiento de datasets de la Dark Web, como se muestra en la Figura 1, presenta una metodología para procesar datasets destinados a su posterior clasificación mediante los LLMs de OpenAI. Este proceso implica transformar y analizar datos extraídos de ficheros .html a .txt. Se efectúa la conversión de formato, se limpia el contenido mediante la eliminación de caracteres no deseados y se filtran los textos en inglés. Posteriormente, se lleva a cabo un análisis gramatical utilizando spaCy para descomponer las frases. Finalmente, se generan nuevos ficheros .txt mediante la selección de frases clave para su uso en análisis posteriores con herramientas avanzadas de procesamiento de lenguaje natural.

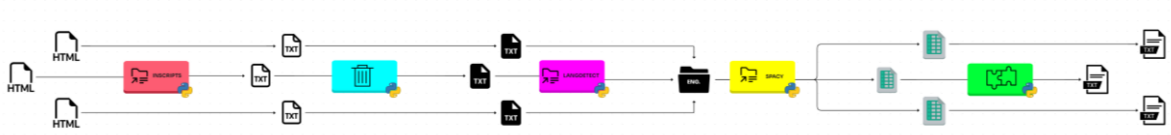

Figura 1. Pasos para el procesamiento de datasets de la Dark Web para su clasificación con la API de OpenAI
Fuente: Departamento de Ciencias de la Computación, Universidad de Alcalá (2024)

3.1.1 Conversión de Ficheros .html a .txt

El primer paso de la metodología implica la transformación de ficheros en formato .html a archivos de texto plano .txt. Esta conversión permite facilitar el procesamiento y análisis de los datos, ya que los ficheros .html contienen etiquetas y elementos que no son relevantes para el análisis textual.

Para realizar esta transformación, se utiliza la biblioteca de Python inscriptis. Esta herramienta extrae el contenido textual de los archivos .html, eliminando las etiquetas y elementos innecesarios, y garantiza que el texto resultante esté listo para su uso en los siguientes pasos del proceso. La conversión se lleva a cabo de forma sistemática, asegurando que cada archivo .html se procese de manera eficiente y que el contenido extraído mantenga la coherencia del texto original.

Al concluir este paso, se obtienen múltiples archivos .txt que contienen el texto extraído de los documentos .html, lo que establece una base para las etapas posteriores de limpieza y análisis.

3.1.2 Limpieza del Contenido de los Ficheros .txt

El segundo paso de la metodología consiste en limpiar el contenido de los archivos .txt generados en el paso anterior. Este proceso es fundamental para asegurar la adecuación de los datos para análisis posteriores y se realiza mediante scripts de Python diseñados para este propósito.

En primer lugar, se eliminan los símbolos repetidos, sustituyéndolos por un único símbolo. También se eliminan los espacios innecesarios y los saltos de línea repetidos. Todo esto se lleva a cabo mediante expresiones regulares para garantizar la consistencia del contenido.

Se procede a detectar y manejar la codificación de los archivos, asegurando una correcta lectura del texto y evitando problemas derivados de codificaciones inadecuadas. Además, se actualiza el contenido para corregir formatos específicos, como la eliminación de múltiples guiones bajos consecutivos.

Finalmente, se dividen los textos en fragmentos más manejables si su longitud es excesiva, lo que facilita el procesamiento en etapas posteriores. Este enfoque permite crear un conjunto de datos limpio y estructurado, lo que resulta esencial para el análisis gramatical y otras tareas de procesamiento de lenguaje natural en los pasos siguientes.

3.1.3 Filtrado de los Ficheros .txt por Idioma

En el tercer paso de la metodología, se lleva a cabo el filtrado de los archivos .txt limpios para seleccionar aquellos que están escritos en inglés. Este proceso utiliza la librería de Python langdetect, que identifica el idioma de cada archivo de texto.

Primero, se aplica la herramienta de detección de idioma a cada uno de los ficheros .txt. Esta herramienta evalúa el contenido textual y determina si está escrito en inglés. Los textos que no cumplen con este criterio son descartados, asegurando que solo se procesen aquellos que se ajustan a los requisitos del análisis posterior.

Durante este proceso, se garantiza que la selección de los textos en inglés se realice de manera eficiente y precisa. Esto es fundamental, ya que el análisis gramatical y el procesamiento posterior se centrarán en un corpus homogéneo que maximiza la relevancia y efectividad de las herramientas de procesamiento de lenguaje natural que se aplicarán en los pasos siguientes. Al finalizar este paso, se obtiene un conjunto de archivos .txt que cumplen con el criterio de idioma, listos para su análisis detallado en la etapa posterior.

3.1.4 Análisis Gramatical de los Ficheros .txt en Inglés

En el cuarto paso de la metodología, se realiza el *POS tagging* (etiquetado de partes del discurso) de los archivos .txt confirmados como escritos en inglés. Para este análisis, se utiliza la librería de Python spaCy, que proporciona herramientas para el procesamiento de lenguaje natural.

Los archivos .txt en inglés se importan al entorno de análisis, y spaCy se emplea para dividir el texto en frases. A la división de frases que realiza spaCy, se añadió la división por saltos de línea, dado que los textos en la Dark Web a menudo carecen de una puntuación correcta, lo que generaba errores en la división de frases realizada únicamente por spaCy. Cada frase se analiza mediante *POS tagging* para identificar componentes gramaticales clave, como sustantivos, verbos, adjetivos, adverbios, pronombres, números y símbolos.

Los resultados del análisis se almacenan en un archivo Excel, creando un conjunto de datos estructurado que incluye detalles gramaticales de cada frase. Este conjunto de datos facilita el acceso y la revisión de la información obtenida. Además, se implementan medidas para manejar archivos grandes sin necesidad de dividirlos en fragmentos más pequeños, asegurando un procesamiento eficiente y evitando la pérdida de información relevante durante el análisis. Al finalizar este paso, se obtiene un conjunto de datos que proporciona una comprensión detallada de la estructura gramatical de los textos en inglés, preparando el terreno para las fases posteriores del estudio.

3.1.5 Composición de Nuevos Ficheros .txt

En el quinto paso de la metodología, se realiza la composición de nuevos archivos .txt a partir de los datos obtenidos en el análisis gramatical realizado en el paso anterior. Este proceso implica la filtración y extracción de frases de los archivos Excel generados en base a criterios específicos.

Se utilizan únicamente las frases que contienen un verbo, un sustantivo o un número acompañado de un símbolo. El resto de frases se descartan, ya que no cumplen con los requisitos mínimos necesarios para el análisis. Este enfoque asegura que solo se seleccionen frases relevantes que aporten valor al procesamiento posterior.

Una vez identificadas las frases que cumplen con los criterios, estas se trasladan a un archivo de texto final. Este archivo contiene exclusivamente el contenido filtrado y estructurado según los parámetros definidos, listo para su uso en tareas posteriores. Al concluir este paso, se generan nuevos archivos .txt, optimizados para ser utilizados en análisis avanzados con herramientas de procesamiento de lenguaje natural, como LLMs. Esta fase finaliza la organización y preparación de los datos para el análisis futuro.

4. RESULTADOS Y DISCUSIÓN

El dataset DUTA-10K se diseñó para el estudio y clasificación de servicios ocultos (Hidden Service, HS) en la red Tor, ofreciendo una base de datos etiquetada para la investigación de actividades sospechosas y normales en dicha red. Contiene 10.367 direcciones únicas de servicios ocultos, organizadas en 25 categorías.

Los servicios ocultos en DUTA-10K se clasifican en tres categorías principales: actividades sospechosas, que constituyen el 20% del conjunto de datos; actividades normales, con un 48%; y un 32% corresponde a dominios no clasificados debido a la inaccesibilidad de su contenido.

4.1. OI1: Optimización del Uso de *Tokens* en Modelos *LLM* de *OpenAI*

Los resultados indican que el proceso de limpieza y optimización de los archivos de texto ha tenido un impacto significativo en la reducción de tokens procesados sobre los documentos que se encontraban en inglés, como se muestra en la Tabla 3. Esto se logró eliminando caracteres innecesarios y normalizando el contenido, lo que resultó en una disminución total del 88% en el número de palabras en los documentos en inglés, es decir, en 8.645 direcciones únicas del dataset DUTA-10K.

En la Tabla 3, se refleja una disminución significativa en el número de palabras en varias categorías del dataset DUTA-10K tras el procesamiento de los datos. Por ejemplo, en la categoría *Marketplace*, el número de palabras fue de 17.050,516 a 519,12, con una reducción porcentual del 96,95%. En la categoría *Casino*, la reducción fue aún mayor, con 51.524,538 palabras menos, lo que supone una disminución del 99,27%.

Estas reducciones impactan directamente en el número de tokens procesados por los modelos de OpenAI, lo que disminuye los costes operativos, dado que OpenAI cobra en función de los tokens utilizados. Esto optimiza los recursos computacionales y permite una clasificación más eficiente de contenidos con los modelos LLM.

4.2 OI2: Análisis de Modelos *LLM* de *OpenAI* para la Clasificación en la *Dark Web*

El análisis de los costes operativos se realizó utilizando una muestra de 1.000 direcciones únicas aleatorias del dataset DUTA-10K, clasificados con el modelo GPT-3.5. Como se muestra en la Tabla 4, el procesamiento de datos resultó en una reducción significativa del número de palabras empleadas, lo que impacta directamente en los costes operativos. En la versión original del dataset, se utilizaron 1.209.164 tokens, mientras que después del procesamiento, esta cifra se redujo a 1.080.141 tokens.

El rendimiento del modelo GPT-3.5 Turbo se mantuvo constante tras la reducción de tokens, sin variaciones significativas en precisión, sensibilidad y F1-score, como se muestra en la Tabla 4. En el conjunto de datos original, la precisión fue del 67,05%, la sensibilidad del 78,61% y el F1-score del 70,29%. Después del procesamiento, los valores fueron similares: precisión del 67,88%, sensibilidad del 76,98% y F1-score del 69,1%. Para los valores ponderados, la precisión fue del 82,75% y 83,03%, la sensibilidad del 78,8% y 79%, y el F1-score del 80,36% y 80,46% antes y después del procesamiento, respectivamente.

Tabla 3. Comparativa de la variación en Palabras en las clases del dataset DUTA-10K sobre los archivos en inglés

Categorías	Total	Palabras previo	Palabras post	Diferencia Palabras	Reducción Palabras
Art	14	1.120,214	544,214	-576	-51,42%
Casino	26	51.905,308	380,769	-51.524,538	-99,27%
Counterfeit Credit-Cards	390	1.716,305	498,064	-1.218,241	-70,98%
Counterfeit Money	81	19.744,889	352,852	-19.392,037	-98,21%
Counterfeit Personal-Identification	60	1.258,45	681,917	-576,533	-45,81%
Cryptocurrency	840	3.251,16	965,788	-2.285,371	-70,29%
Cryptolocker	50	115,42	98,84	-16,58	-14,36%
Down	754	613,046	59,958	-553,089	-90,22%
Drugs	290	5.406,141	395,248	-5.010,893	-92,69%
Empty	1351	225,331	12,024	-213,307	-94,66%
Forum	128	5.932,547	375,922	-5.556,625	-93,66%
Fraud	19	367,474	309	-58,474	-15,91%
Hacking	182	3.749,945	768,451	-2.981,495	-79,51%
Hosting	2223	2.858,93	755,78	-2.103,151	-73,56%
Human-Trafficking	3	168	60,667	-107,333	-63,89%
Leaked-Data	17	23.255,059	15.229,353	-8.025,706	-34,51%
Library	30	18.022,7	602,7	-1.7420	-96,66%
Locked	680	199,235	45,104	-154,131	-77,36%
Marketplace	189	17.050,516	519,12	-1.6531,396	-96,95%
Personal	415	6.503,93	6.44,759	-5.859,171	-90,09%
Politics	2	4.559,5	4050	-509,5	-11,17%
Porno	225	19.948,196	956,991	-18.991,204	-95,20%
Religion	16	792,563	208	-584,563	-73,76%
Services	284	2.131,746	558,468	-1.573,278	-73,80%
Social-Network	290	10.314,266	1.144,872	-9.169,393	-88,90%
Violence	86	1.178,718	378,306	-800,412	-67,91%
TOTAL	**8.645**	**101.196,294**	**15.300,084**	**-85.896,211**	**-84,88%**

Estos resultados indican que el procesamiento y la optimización del uso de tokens lograron reducir los costes operativos sin afectar el rendimiento del modelo en términos de precisión y sensibilidad. La disminución en el número de tokens procesados implicó una reducción en los recursos computacionales y, por tanto, en los costes asociados, sin comprometer la calidad de la clasificación.

Tabla 4. Comparación del rendimiento de la clasificación en el dataset DUTA-10K entre contenidos originales y contenidos optimizados

Muestra aleatoria con GPT-3.5 Turbo		Precisión	Sensibilidad	F1
Media	**DUTA Original**	67,05%	78,61%	70,29%
	DUTA Optimizado	67,88%	76,98%	69,1%
Media ponderada	**DUTA Original**	82,75%	78,8%	80,36%
	DUTA Optimizado	83,03%	79%	80,46%

5. CONCLUSIONES

Este estudio ha examinado el procesamiento de datasets de la Dark Web utilizando modelos LLM de OpenAI, con el fin de optimizar el uso de tokens y disminuir los costos operativos en la clasificación de contenidos. Gracias a la metodología implementada, se logró una reducción significativa en el número de caracteres y palabras procesados, lo que resultó en una disminución del 85% en el consumo de tokens de las direcciones con contenido en inglés. Esta optimización permitió mantener la eficiencia del modelo GPT-3.5 Turbo en términos de precisión, sensibilidad y F1-score, sin afectar el rendimiento en la clasificación.

La optimización del uso de tokens se ha mostrado como una estrategia efectiva para reducir los recursos computacionales y los costes operativos, especialmente en el análisis de grandes volúmenes de datos, como los contenidos de la Dark Web. Aunque el rendimiento del modelo se mantuvo constante, la aplicación de técnicas de procesamiento previo al análisis ofrece beneficios económicos significativos, permitiendo la implementación de estos modelos en entornos prácticos.

AGRADECIMIENTOS

La publicación es parte del Proyecto PARCHE, con referencia PID2021-125645OB-I00 financiado por MCIN/AEI/10.13039/501100011033/FEDER, UE.

REFERENCIAS

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D.R., Smith, N.A., Tsvetkov, Y. (2023). Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. https://doi.org/10.48550/ARXIV.2305.13707

Al Nabki, M.W., Fidalgo, E., Alegre, E., de Paz, I. (2017). Classifying Illegal Activities on Tor Network Based on Web Textual Contents, in: Lapata, M., Blunsom, P., Koller, A. (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Presented at the EACL 2017, Association for Computational Linguistics, Valencia, Spain, pp. 35–43.

Al-Nabki, M.W., Fidalgo, E., Alegre, E., Fernández-Robles, L. (2019). ToRank: Identifying the most influential suspicious domains in the Tor network. Expert Syst. Appl. 123, 212–226. https://doi.org/10.1016/j.eswa.2019.01.029

Avarikioti, G., Brunner, R., Kiayias, A., Wattenhofer, R., Zindros, D. (2018). Structure and Content of the Visible Darknet. https://doi.org/10.48550/arXiv.1811.01348

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., Brightwell, T. (2023). Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification, in: Métais, E., Meziane, F., Sugumaran, V., Manning, W., Reiff-Marganiec, S. (Eds.), Natural Language Processing and Information Systems, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp. 3–17. https://doi.org/10.1007/978-3-031-35320-8_1

Jin, Y., Jang, E., Lee, Y., Shin, S., Chung, J.-W. (2022). Shedding New Light on the Language of the Dark Web. https://doi.org/10.48550/arXiv.2204.06885

Kalyan, K.S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat. Lang. Process. J. 6, 100048. https://doi.org/10.1016/j.nlp.2023.100048

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit, in: Bontcheva, K., Zhu, J. (Eds.), Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland, pp. 55–60. https://doi.org/10.3115/v1/P14-5010

OpenAI Platform [WWW Document], n.d. URL https://platform.openai.com (accessed 2.29.24).

Prado Sánchez, V.P., Domínguez Díaz, A., Marcos, L., Martínez Herráiz, J.J. (2024). Clasificación zero-shot de contenidos de la Dark Web mediante GPT-3.5: Evaluación de rendimiento y análisis de errores del clasificador. Universidad de Sevilla. Escuela Técnica Superior de Ingeniería Informática.

Prado-Sánchez, V.-P., Domínguez-Díaz, A., Rodríguez, D., Martínez, J.-J., n.d. How can ChatGPT help humans in Dark Web content classification? Assessing GPT models reliability and effects of explanations on human decisions.

Roumeliotis, K.I., Tselikas, N.D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. Future Internet 15, 192. https://doi.org/10.3390/fi15060192

Singh, A.K., Strouse, D.J. (2024). Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs. https://doi.org/10.48550/arXiv.2402.14903