# Clasificación zero-shot de contenidos de la Dark Web mediante GPT-3.5: Evaluación de rendimiento y análisis de errores del clasificador

Víctor-Pablo Prado-Sánchez, Adrián Domínguez-Díaz, Luis de-Marcos, José-Javier Martínez-Herráiz
Departamento de Ciencias de la Computación, Universidad de Alcalá
Alcalá de Henares, España
victor.prado@uah.es, adrian.dominguez@uah.es, luis.demarcos@uah.es, josej.martinez@uah.es

*Resumen*- **La clasificación automática de contenidos de la Dark Web es relevante para la detección e investigación de actividades delictivas. Sin embargo, la escasez de datos etiquetados impone límites al uso de clasificadores supervisados. Los grandes modelos de lenguaje, con capacidad de clasificación en categorías en las que no han sido entrenados, no están sujetos a estas limitaciones. El estudio se centra en evaluar el rendimiento del modelo de lenguaje GPT-3.5 de OpenAI para clasificar contenido de texto de la Dark Web bajo un enfoque de *zero-shot prompting*. El modelo alcanza un valor F1 ponderado del 80,5%, detectándose grandes diferencias entre categorías y distintas problemáticas que limitan su rendimiento respecto al estado del arte en clasificadores supervisados. El análisis de los errores cometidos permite identificar contextos de aplicación en los que podría resultar competitivo bajo las condiciones de estudio, así como sugerir distintas estrategias para mejorar su rendimiento.**

***Index Terms*- Dark Web, Darknet, Ciberseguridad, Modelos de lenguaje, LLM, NLP, Zero-shot learning, Zero-shot prompting, ChatGPT, GPT-3.5, Prompt engineering, Overfitting.**

**Tipo de contribución:** *Investigación original*

## I. Introducción

El mayor volumen y complejidad de actividades desarrolladas en la Dark Web plantea desafíos para la clasificación de sus contenidos, especialmente en el contexto de actividades ilegales y potencialmente dañinas [1]. Por ello, se ha prestado una creciente atención a la aplicación de técnicas de aprendizaje automático para abordar esta problemática [2]. En esta línea, se pueden encontrar diversos enfoques basados en modelos de *machine learning* entrenados mediante aprendizaje supervisado para detectar y clasificar actividades ilegales en la Dark Web [3].

Dada la naturaleza lingüística de los contenidos a clasificar, también se ha explorado el desempeño en esta tarea de los modelos de lenguaje pre-entrenados. Por ejemplo, el uso de modelos de la familia BERT (Bidirectional Encoder Representations from Transformers) [4] ha resultado ser muy efectivo para la clasificación de texto en diversos dominios [5], incluido el de contenidos de sitios web de la Dark Web [3], [6]. La necesidad de evaluar el rendimiento de distintos modelos en este dominio ha dado lugar a la creación de conjuntos de datos con textos de sitios de la Dark Web manualmente etiquetados, como DUTA (Darknet Usage Text Addresses) [2] y CoDA (Comprehensive Darkweb Annotations) [3].

Pese al buen rendimiento alcanzado, los modelos mencionados requieren de una fase previa de aprendizaje supervisado o fine-tunning del modelo de lenguaje sobre una parte del conjunto de datos para aprender a clasificarlos de forma efectiva [5]. Esto supone que el rendimiento de los clasificadores en entornos reales estaría limitado por la disponibilidad y calidad de datos etiquetados, así como por la posibilidad de sufrir *overfit* a los datos de entrenamiento, teniendo dificultades para adaptarse a la naturaleza dinámica y heterogénea de la Dark Web [2].

Para solventar esta limitación algunos estudios proponen el uso de aprendizaje activo, en el que tras una fase de entrenamiento supervisado, un clasificador humano continuaría refinando el modelo de forma iterativa mediante etiquetado manual de los documentos en los que el modelo detecte mayor incertidumbre [7]. Esta estrategia se ha demostrado efectiva para mejorar el rendimiento en la clasificación de contenidos de la Dark Web [1], para detectar contenido malicioso en redes sociales [8], así como para otros dominios de aplicación basados en la clasificación de contenido de texto [9]. Aunque en menor medida que un aprendizaje puramente supervisado, el aprendizaje activo sigue requiriendo de un esfuerzo significativo por parte de clasificadores humanos.

Una alternativa que minimizaría la necesidad de datos etiquetados y el esfuerzo de clasificadores humanos es el uso de modelos de lenguaje de gran tamaño, como son los modelos GPT-3 y sus derivados [10], [11], mediante *zero-shot o few-shot prompting*. Este modo de funcionamiento se basa en aprovechar las capacidades lingüísticas de los LLMs para clasificar textos en categorías en las que no ha sido entrenado, proporcionándole tan solo una descripción de las mismas a través del *prompt* (*zero-shot*), o acompañándolas de algunos ejemplos (*few-shot*) [12]. Esto enfoque ha sido probado para la clasificación de textos financieros, ofreciendo resultados rápidos y precisos [13]. Del mismo modo, se ha identificado el potencial de ChatGPT para detectar contenido perjudicial en redes sociales con una precisión del 80% [14]. Sin embargo, se han planteado interrogantes sobre la fiabilidad de los modelos GPT en este tipo de tareas debido a su naturaleza no determinista [12].

Bajo el enfoque *zero-shot* o *few-shot*, adquiere la máxima importancia el *prompt* con que se alimenta al modelo de lenguaje. El *prompt engineering* estudia distintas técnicas sobre cómo estructurar el texto de entrada para los LLMs con el fin de optimizar su eficacia. Una aplicación adecuada del *prompt engineering* puede desbloquear las capacidades de los LLMs y mitigar la tendencia a la alucinación de las máquinas [15]. Se ha observado que la redacción del *prompt* es un factor

crítico para obtener el razonamiento adecuado en el modelo, y que aspectos aparentemente menores del mismo afectan significativamente su rendimiento [16].

Tras la revisión del estado del arte, no se han encontrado por nuestra parte estudios que analicen el rendimiento de LLMs para la clasificación de contenidos de la Dark Web bajo un modelo de *zero-shot prompting*. En este estudio se propone evaluar la eficacia del modelo de lenguaje GPT-3.5 de OpenAI para clasificar el conjunto de datos CoDA bajo un modelo *zero-shot* y un *prompt* mínimo que incluye la descripción de las clases. Los objetivos de la investigación son los siguientes:

- OI1: Evaluar el rendimiento del modelo de lenguaje en la clasificación e contenidos de texto de la Dark Web bajo un enfoque *zero-shot prompting*
- OI2: Identificar los errores más habitualmente cometidos por el clasificador y las causas que provocan dichos errores.
- OI3: Determinar los supuestos bajo los que este modelo puede ofrecer resultados competitivos en entornos de producción en comparación con clasificadores supervisados y semi-supervisados.

Este trabajo contribuye a un mayor conocimiento sobre las herramientas disponibles para el análisis de actividad criminal en la Dark Web, mostrando el rendimiento en tareas de clasificación del modelo GPT-3.5 mediante *zero-shot prompting*, un método disponible para el público general y caracterizado por ofrecer potentes capacidades de procesamiento de lenguaje natural con un mínimo esfuerzo y complejidad técnica. A través del análisis detallado de los errores más frecuentes cometidos por el clasificador, el trabajo ayuda a identificar en qué tareas podría usarse este modelo de forma directa, y en qué otras tareas sería necesario continuar investigando para que el rendimiento de este método pueda acercarse al de clasificadores entrenados mediante aprendizaje supervisado o semi-supervisado.

El resto de este documento se estructura de la siguiente manera. En la Sección II se proporciona una visión general de los trabajos relacionados sobre clasificación de contenidos de texto bajo distintos algoritmos de *machine learning* y modelos de lenguaje. La Sección III se detalla la metodología empleada para evaluar el desempeño del modelo GPT-3.5 en la clasificación del conjunto de datos CoDA. La Sección IV presenta los resultados, profundizando en los errores de precisión y sensibilidad más frecuentes. En la Sección V se discuten los resultados obtenidos en base a los objetivos de investigación. Finalmente, en la Sección VI se exponen las conclusiones y se discuten las líneas de trabajo futuro.

## II. Estado del arte

### A. Clasificación mediante modelos supervisados

En el contexto de la investigación sobre la Dark Web y la detección de actividades ilegales en línea, se han desarrollado varios estudios que evalúan diferentes algoritmos de machine learning y modelos de lenguaje en la clasificación de contenidos de texto bajo un modelo de aprendizaje supervisado.

En [17] se aplica el algoritmo KNN (K-Nearest Neighbor) para categorizar el contenido de bloques de texto en páginas web ocultas en idioma ruso, obteniendo grandes diferencias de precisión según las categorías, con resultados muy positivos en detección de drogas o páginas para adultos y mucho peores en páginas de venta de documentos falsificados o medios de pago. En [18] se integra TF-IDF (Term Frequency-Inverse Document Frequency), PCA (Principal component analysis) y SVM (Support Vector Machine), para categorizar productos publicados en mercados anónimos, específicamente en el mercado Agora, consiguiendo una precisión del 79% al clasificar en 12 categorías.

Algunos de estos estudios, además de evaluar distintos algoritmos de clasificación, también hacen públicos conjuntos de datos de la Dark Web facilitando la replicabilidad y el desarrollo de nuevos estudios comparativos.

En [2] se presenta el conjunto de datos DUTA para dominios activos en Dark Web, con 6.831 documentos clasificados en 26 categorías. Se compararon diversas técnicas de procesamiento de lenguaje natural y modelos de machine learning para categorizar estos documentos, encontrando que la combinación de TF-IDF y Regresión Logística logra un valor F1 ponderado del 97% en la clasificación de un subconjunto del dataset en 9 categorías.

En [3] se introduce el conjunto de datos CoDA, que consta de 10,000 documentos web obtenidos de la Dark Web para análisis de texto. Se examinan las diferencias lingüísticas entre la Dark Web y la Surface Web y se analizan las diferencias entre los conjuntos de datos CoDA y DUTA [2]. También se evalúa el rendimiento de varios métodos de clasificación de páginas de la Dark Web alcanzando un valor F1 ponderado de 92.49% mediante el uso del modelo de lenguaje pre-entrenado BERT, tras un proceso de fine-tunning para entrenarle en la clasificación de documentos. En [6] se presenta un clasificador mejorado basado en una adaptación de BERT pre-entrenada con textos obtenidos de la Dark Web, con el que se alcanza un valor F1 de 94.25% en la clasificación de los contenidos del conjunto CoDA.

### B. Clasificación mediante modelos semi-supervisados

El aprendizaje semi-supervisado y el aprendizaje activo también han sido evaluados en la clasificación de contenidos de texto en casos donde no se disponía de un gran dataset y se quería minimizar el esfuerzo humano para entrenar un modelo de clasificación.

En [9] se introduce DUALIST, un método y herramienta de aprendizaje activo sobre documentos de texto que solicita y aprende etiquetas mediante interacción con un humano. Este método demuestra ser efectivo para distintos fines como la clasificación de contenidos o el análisis de sentimientos.

Enfocado en la detección de contenido malicioso en foros, [8] presenta un algoritmo de aprendizaje activo, poniéndolo a prueba en distintos contextos como contenido terrorista, fake news o tráfico humano. Se consiguen valores F1 competitivos habiendo entrenado sobre conjuntos de datos muy reducidos respecto a los modelos supervisados.

En el contexto del tráfico de personas, [19] presenta un modelo capaz de clasificar anuncios con posibles vínculos con el tráfico humano. Se utiliza un enfoque de aprendizaje semi-supervisado combinando un pequeño conjunto de datos etiquetados por un experto con otros no etiquetados. El clasificador binario alcanza un valor AUC de 0.91, resultando el enfoque muy efectivo en clasificación binaria con un dataset

acotado.

Ya centrado en la Dark Web, [1] aplica aprendizaje activo para mejorar el desempeño de un clasificador supervisado entrenado sobre un pequeño conjunto de 200 documentos. Este clasificador alcanza un valor F1 del 60% en clasificación en diez categorías de contenido, así como del 85% en clasificación binaria indicando si los contenidos son legales o ilegales. En línea con el estudio anterior, los resultados demuestran ser especialmente efectivos en clasificación binaria.

### C. *Clasificación mediante zero-shot y few-shot prompting*

Las capacidades de procesamiento de lenguaje natural de los LLMs más recientes como son los modelos GPT permiten abordar tareas de clasificación de texto sin la necesidad de un entrenamiento previo específico para las categorías de clasificación.

En el ámbito financiero, [13] propone el uso de modelos conversacionales GPT, como GPT-3.5 y GPT-4, para realizar clasificación de textos financieros con pocos ejemplos disponibles y comparar su rendimiento con otros modelos de lenguaje pre-entrenados y ajustados con *fine-tunning*. GPT-3.5 y GPT-4 superan al resto de modelos en clasificación con 1 o 3 ejemplos, con valores F1 de 75,2% y 83,1% respectivamente, pero quedan atrás respecto a modelos de lenguaje con *fine-tunning*. Se destaca la facilidad para implementar clasificadores efectivos con un mínimo esfuerzo técnico mediante los servicios ofrecidos por empresas como OpenAI, aunque se señala como punto negativo el elevado coste que puede llegar a tener para organizaciones pequeñas.

En relación con la detección de contenido perjudicial en redes sociales, [14] aborda este desafío mediante el uso de ChatGPT como modelo de detección. Se comparan sus resultados con anotaciones humanas, demostrando una precisión del 80% en la detección de contenido perjudicial. Se destaca la consistencia de las clasificaciones, aunque se reconoce la influencia del *prompt* utilizado en su rendimiento.

Con un enfoque más general, [12] examina la fiabilidad de ChatGPT para tareas de anotación y clasificación de texto en diversos contextos de aplicación. Si bien se reconoce su potencial, se plantean preocupaciones sobre su naturaleza no determinista, que puede producir resultados variables incluso ante pequeños cambios en las entradas. Se sugiere precaución al utilizar ChatGPT para estas tareas sin una validación adicional, como comparaciones con datos anotados por humanos.

Finalmente, en [20] se evalúa ChatGPT (GPT-3.5) y el modelo GPT 4 en diversas tareas de procesamiento de lenguaje natural bajo un enfoque *zero-shot* o *few-shot*. Se alcanza la conclusión de que los modelos GPT pueden tener un buen desempeño, mejor en tareas objetivas que subjetivas, pero siempre por debajo de las técnicas específicas más avanzadas, con una perdida de calidad mayor cuanto más compleja es la tarea.

## III. METODOLOGÍA

### A. *Selección del conjunto de datos CoDA*

Se empleó el conjunto de datos CoDA (*Comprehensive Darkweb Annotations*), el cual representa un recurso público valioso para el análisis de la Dark Web, comprendiendo una colección de 10,000 documentos web destinados a la investigación basada en texto en este entorno.

Estos documentos abarcan una amplia gama de temas y se clasificaron en diez categorías temáticas distintas. Principalmente en inglés, los documentos se han obtenido de servicios *onion* en Tor, es decir, en la Dark Web, lo que proporciona una idea significativa de este espacio digital poco explorado. Los documentos fueron clasificados en diez categorías según su temática; donde se incluyen *Drugs*, *Financial*, *Gambling*, *Cryptocurrency (Crypto)*, *Hacking*, *Arms/Weapons (Arms)*, *Violence*, *Electronics*, así como *Pornography* y *Others*. Esta amplia variedad de categorías permite una evaluación de diversos aspectos de la actividad en la Dark Web.

Es necesario destacar, sin embargo, que fue necesario excluir los documentos web categorizados como *Pornography*, al entrar en conflicto con las políticas de contenidos de OpenAI [21]. Estas políticas son aplicables al uso de cualquier servicio de OpenAI, incluyendo ChatGPT, labs.openai.com y la API de OpenAI. Según estas normativas, se prohíbe la construcción de herramientas que puedan ser inapropiadas para menores, lo cual incluye contenido sexualmente explícito o sugerente, a menos que esté creado con fines científicos o educativos. Mediante distintas pruebas se observó que la API implementa controles que, al detectar este tipo de contenidos, devuelven mensajes de error indicando la necesidad de cesar la actividad bajo riesgo de clausura de la cuenta de usuario, lo que hizo inviable el procesamiento de este tipo de contenidos.

### B. *Modelo de OpenAI GPT-3.5*

Para la clasificación de los documentos provenientes de la Dark Web, se decidió emplear el modelo de lenguaje GPT-3.5 de OpenAI, en particular la variante denominada GPT-3.5 Turbo. Esta elección se basó en la capacidad demostrada por este modelo para procesar texto de manera eficiente y comprender el contexto complejo y variado característico de la Dark Web [22].

El último modelo GPT-3.5 Turbo ha sido actualizado para mejorar su precisión al responder en diferentes formatos, además de solucionar un error que afectaba la codificación de texto en llamadas de funciones en idiomas no ingleses. Esta nueva versión permite la generación de hasta 4,096 tokens de salida y amplía la ventana de contexto a 16,385 tokens. Sus datos de entrenamiento están actualizados hasta septiembre de 2021 [23].

La selección de esta variante del modelo se justificó por su mejor relación entre capacidades y coste en comparación con otros modelos como GPT 4, 20 veces más caro en el momento de realizar el estudio, y una mejor adecuación a las necesidades del problema, con un contexto de entrada de 16.385 tokens que resulta adecuado para el tamaño de documentos a clasificar.

### C. *Configuración del modelo*

El modelo recibió los datos necesarios pare realizar la tarea de clasificación bajo un modelo de *zero-shot prompting* [15]. El *prompt* en los modelos GPT de OpenAI implica la definición de dos roles principales: *system* y *user*. El rol *system* actúa proporcionando instrucciones de alto nivel al modelo, estableciendo así el marco y los criterios para la clasificación de documentos. Por otro lado, el rol *user* permite presentar consultas o indicaciones específicas relacionadas con el contenido de los documentos que se van a clasificar. Estos elementos son esenciales para guiar las respuestas generadas por los modelos GPT [24], [25].

Siguiendo el modelo *zero-shot prompting*, bajo el rol *system*, se suministraron al modelo descripciones detalladas de cada clase, según fueron definidas en el protocolo de clasificación de documentos con el que se elaboró el conjunto de datos CoDA [3]. Posteriormente, bajo el rol *user*, se procedió a ingresar el contenido de los documentos web individuales de CoDA para su análisis y clasificación. Este paso implicó proporcionar al modelo el texto de cada documento y solicitar su clasificación conforme a los criterios previamente definidos en el rol de *system*. El lenguaje utilizado para el *prompt* fue inglés, dado que el grueso de documentos incluidos en CoDA está en este idioma. Los parámetros de la llamada al modelo GPT fueron los que el sistema establece por defecto.

### D. Evaluación del rendimiento del clasificador

Se evaluó el desempeño del clasificador lanzando una petición individual por cada documento del conjunto de datos y almacenando la respuesta proporcionada por ChatGPT. Antes de calcular las métricas de evaluación del modelo, se realizó una revisión y corrección manual de las respuestas proporcionadas, ya que, dada su naturaleza generativa y no determinista, un pequeño porcentaje de ellas incluían pequeñas diferencias respecto a los valores de salida esperados.

Posteriormente se calcularon las métricas de precisión, sensibilidad y F1 para cada categoría y las medias macro y ponderada. Mediante las matrices de confusión, se calcularon los errores de falso positivo y falso negativos más frecuentemente cometidos por el clasificador según la categoría erróneamente asignada. Para cada categoría $cat$ en la que se haya cometido un falso positivo o negativo con otra categoría $err$, se calculó la frecuencia del error con las siguientes fórmulas:

$$\%FP_{cat\_err} = FP_{cat\_err} / (TP_{cat} + FP_{cat})$$

$$\%FN_{cat\_err} = FN_{cat\_err} / (TP_{cat} + FN_{cat})$$

Siendo $FP_{cat\_err}$ el número de documentos de la categoría $err$ incorrectamente clasificados como $cat$, $FN_{cat\_err}$ el número de documentos de la categoría $cat$ incorrectamente clasificados como $err$, y $TP_{cat}$, $FP_{cat}$ y $FN_{cat}$ el número total de verdaderos positivos, falsos positivos y falsos negativos de la categoría $cat$ respectivamente.

Tabla I
RENDIMIENTO CLASIFICACIÓN GPT-3.5 TURBO
CON CATEGORÍA OTHERS INCLUIDA

| Categoría | Precisión | Sensibilidad | F1 |
|---|---|---|---|
| Gambling | 98.0% | 88.3% | 92.9% |
| Arms | 96.6% | 77.0% | 85.7% |
| Drugs | 93.9% | 94.9% | 94.4% |
| Violence | 87.2% | 65.8% | 75.0% |
| Others | 85.0% | 66.0% | 74.3% |
| Financial | 80.0% | 89.5% | 84.5% |
| Electronics | 77.9% | 76.8% | 77.3% |
| Crypto | 62.6% | 89.6% | 73.7% |
| Hacking | 58.7% | 91.5% | 71.5% |
| Media | 82.2% | 82.2% | 81.0% |
| Media Pond. | 83.4% | 79.7% | 80.5% |

Tabla II
DETALLE DE ERRORES DE FALSO POSITIVO
MÁS FRECUENTES POR CATEGORÍA (> 1%)
CON CATEGORÍA OTHERS INCLUIDA

| Categoría | Error #1 | Error #2 | Error #3 |
|---|---|---|---|
| Arms | Violence (1.5%) | | |
| Crypto | Others (28.0%) | Financial (4.3%) | Drugs (2.2%) |
| Drugs | Others (4.8%) | | |
| Electronics | Others (19.0%) | | |
| Financial | Others (10.2%) | Electronics (6.1%) | Hacking (1.6%) |
| Gambling | Others (1.0%) | | |
| Hacking | Others (29.6%) | Crypto (3.1%) | Arms (2.7%) |
| Others | Violence (5.6%) | Arms (3.1%) | Gambling (2.8%) |

## IV. RESULTADOS

### A. Rendimiento del clasificador

En la *T*abla I se presentan los resultados de rendimiento de la clasificación realizada por GPT-3.5 Turbo sobre el dataset CoDA, exceptuando los documentos clasificados como pornografía. Con una precisión ponderada del 83,4%, los valores de precisión más altos se obtienen en las categorías *Gambling* (98%) y *Arms/Weapons* (96,7%), mientras que en los más bajos se encuentran las categorías *Cryptocurrencies* (62,6%) y *Hacking* (58,7%), mostrando una gran diferencia en la precisión del clasificador entre distintas categorías.

En la Tabla II se presenta un análisis detallado de los errores de precisión cometidos, con los tres errores de falso positivo más frecuentes (min. >1%) en cada categoría. Se observa que los errores más frecuentes son los falsos positivos con documentos pertenecientes a la categoría *Others*, con porcentajes especialmente altos en *Hacking* (29,6%), *Cryptocurrency* (28%) y *Electronics* (19%).

### B. Rendimiento con categoría Others excluida

Dados los problemas a la hora de clasificar correctamente los documentos en la categoría *Others*, discutidos más adelante, se volvió a procesar el conjunto de datos CoDA excluyendo todos los documentos originalmente asignados a dicha categoría y eliminándola del *prompt*. A cambio, se incluyó un mensaje invitando al modelo a elegir prioritariamente alguna de las restantes categorías, o, en

Tabla III
RENDIMIENTO CLASIFICACIÓN GPT 3.5 TURBO
CON CATEGORÍA OTHERS EXCLUIDA

| Categoría | Precisión | Sensibilidad | F1 |
|---|---|---|---|
| Gambling | 99.0% | 90.2% | 94.4% |
| Drugs | 98.7% | 95.0% | 96.8% |
| Arms | 98.0% | 81.5% | 89.0% |
| Violence | 91.8% | 80.6% | 85.8% |
| Financial | 87.8% | 90.2% | 89.0% |
| Crypto | 85.2% | 90.5% | 87.8% |
| Electronics | 83.3% | 77.2% | 80.1% |
| Hacking | 77.0% | 93.7% | 84.5% |
| Media | 91.0% | 89.0% | 89.7% |
| Media Pond. | 91.0% | 89.0% | 89.7% |

Tabla IV
DETALLE DE ERRORES DE FALSO POSITIVO
MÁS FRECUENTES POR CATEGORÍA (> 1%)
CON CATEGORÍA OTHERS EXCLUIDA

| Categoría | Error #1 | Error #2 | Error #3 |
|---|---|---|---|
| Arms | Violence (1.4%) | | |
| Crypto | Financial (6.8%) | Drugs (3.6%) | Electronics (1.7%) |
| Drugs | | | |
| Electronics | Gambling (7.6%) | Violence (2.3%) | Crypto (1.8%) |
| Financial | Electronics (6.6%) | Crypto (1.8%) | Hacking (1.7%) |
| Gambling | | | |
| Hacking | Arms (5.3%) | Violence (5.2%) | Gambling (4.2%) |
| Violence | Arms (7.7%) | | |

Tabla V
DETALLE DE ERRORES DE FALSO NEGATIVO
MÁS FRECUENTES POR CATEGORÍA (> 1%)
CON CATEGORÍA OTHERS EXCLUIDA

| Categoría | Error #1 | Error #2 | Error #3 |
|---|---|---|---|
| Financial | Crypto (5.5%) | Hacking (2.5%) | |
| Crypto | Hacking (4.1%) | Financial (2.5%) | |
| Arms | Hacking (7.0%) | Violence (5.5%) | Electronics (1.2%) |
| Hacking | Financial (2.8%) | Crypto (1.8%) | |
| Drugs | Crypto (2.5%) | | |
| Electronics | Financial (16.0%) | Crypto (3.3%) | |
| Violence | Hacking (8.5%) | Electronics (1.9%) | Arms (1.4%) |
| Gambling | Hacking (4.2%) | Electronics (3.8%) | |

última instancia, a proporcionar un término libre cuando considerase que ninguna de las categorías proporcionadas era adecuada para clasificar el contenido.

En la Tabla III se presentan los resultados de esta clasificación. En este caso, el clasificador consigue una precisión ponderada del 91,0%, con una mejora de 6,6 puntos y la sensibilidad ponderada del 89%, con una mejora de 9,3 puntos, lo que demuestra el impacto de los errores de clasificación con *Others*. A nivel de categoría, se observan precisiones por debajo del 90% en las categorías *Hacking*, *Electronics*, *Cryptocurrency* y *Financial*, así como sensibilidades por debajo del 82% en las categorías *Arms/Weapons*, *Violence* y *Electronics*.

Para profundizar en los errores del clasificador, en las Tablas IV y V se muestra el detalle de los errores de positivo y falso negativo cometidos, indicando los tres errores más frecuentes (min. >1%) en cada categoría. Respecto a los falsos positivos (Tabla IV), se observa un cruce de errores entre las categorías *Arms/Weapons* y *Violence*, ciertos errores especialmente frecuentes (>5%) entre *Cryptocurrency* y *Financial*, *Electronics* y *Gambling*, así como entre *Financial* y *Electronics*, así como una tendencia a clasificar documentos de diversas categorías como *Hacking*. Al revisar los falsos negativos (Tabla V) se observa claramente cómo numerosos documentos de distintas categorías se clasifican incorrectamente como *Hacking*, y en menor medida, como *Cryptocurrency* o *Financial*. Llama la atención un elevado porcentaje de documentos de *Electronics* incorrectamente clasificados como *Financial* (16%).

## V. Discusión

### *OI1 – Rendimiento del clasificador*

Las métricas de evaluación del modelo GPT 3.5 Turbo bajo un enfoque *zero-shot* (Tabla 1) muestran que este es capaz de clasificar con una alta precisión y sensibilidad (F1 > 90%) documentos de la Dark Web relacionados con venta de drogas y apuestas, y, en menor medida (F1 > 80%), relacionados con la venta o fabricación de armas, así como con la falsificación o robo de información financiera. Sin embargo, tiene dificultad para clasificar sitios web dedicados a la venta de productos o servicios de hacking, criptomonedas, dispositivos electrónicos y violencia, presentando bajos valores de precisión o sensibilidad según el caso. Estos resultados sugieren que los modelos GPT 3.5 bajo una modalidad *zero-shot* podrían usarse de forma efectiva como clasificadores binarios en las clases donde es capaz de discriminar de forma efectiva. Sin embargo, su efectividad como clasificador en múltiples clases queda en entredicho por distintos problemas que se discuten a continuación.

### *OI2 - Errores cometidos por el clasificador*

Los resultados de rendimiento sobre el conjunto de datos completo (Tabla I) permiten ver las dificultades para clasificar correctamente documentos pertenecientes a la clase *Others*, definida como la negación de todas las restantes clases. Los falsos positivos más frecuentes (Tabla II) permiten ver que muchos documentos de esta clase se están asignando a otras categorías, con especial frecuencia a *Cryptocurrency*, *Hacking* o *Financial*. Una revisión al texto de algunos de estos documentos incorrectamente clasificados permite ver que frecuentemente incluyen términos que pueden provocar este tipo de error. Algunos ejemplos serían las páginas de enlaces a sitios de la Dark Web, que incluyen breves descripciones y términos relativos al contenido de los sitios enlazados, propios de otras categorías, o páginas de venta de servicios o productos legales que soportan pago con bitcoins e incluyen determinada terminología propia de criptomonedas. La dificultad para clasificar documentos en una clase *Others* también ha sido detectada en clasificadores semi-supervisados que generan una descripción de baja dimensionalidad de la clase [18], mientras que no se observa en clasificadores supervisados [2], [3]. Esto lleva a pensar que, en un contexto de *zero-shot* learning, la descripción de la clase *Others* proporcionada al modelo no resultaría suficiente para clasificar adecuadamente los documentos en esta categoría. Ante la dificultad para describir mejor una clase que se caracteriza por poder dar cabida a una amplia variedad de documentos, surge la duda de si proporcionar a través del *prompt* un protocolo de clasificación detallado, similar al que se proporcionó a los clasificadores humanos durante la elaboración del conjunto de datos, paliaría esta problemática sin necesidad de abandonar el enfoque *zero-shot*.

Cuando se omiten los documentos de la categoría *Others* (Tabla III) observamos que el rendimiento del clasificador aumenta, pero sigue presentando grandes variaciones en la precisión y en la sensibilidad según las categorías. El análisis de falsos positivos en este caso (Tabla IV) muestra las dificultades del modelo de lenguaje para elegir la correcta entre varias categorías cuando sus temáticas tienen cierta superposición, como el porcentaje de documentos clasificados como *Cryptocurrency* pertenecientes a la categoría *Financial*

(6,8%), o clasificados como *Violence* pertenecientes a la categoría *Arms/Weapons* (7,7%). También se intuye una dificultad para clasificar documentos cuando una clase suele incluir términos de otra. Un caso serían los documentos de *Gambling* clasificados como *Electronics*, que probablemente se deban al uso de términos relativos a los dispositivos digitales desde los que realizar apuestas.

El análisis de falsos negativos (Tabla V) también muestra un especial problemática con la clase *Hacking*, y en menor medida *Cryptocurrency* y *Financial*, que se asignan con frecuencia a documentos de distintas categorías. Este problema se ha detectado también en la clasificación supervisada [2], mostrando cómo ciertos términos de estas categorías son ubicuos en la Dark Web. Un ejemplo serían los sitios que, independientemente del producto o servicio a la venta, ofrecen formas de pago mediante tarjeta de crédito o criptomonedas. En el caso de la clase *Hacking*, una explicación adicional es que los clasificadores no dispongan de la comprensión necesaria para distinguir un medio, como el hacking, de un fin, como puede ser el robo de datos financieros, criptomonedas o dispositivos electrónicos. Se plantea la pregunta de si modelos de lenguaje más avanzados, con un *prompt* adecuado, serían capaces de hacer esta diferenciación.

*OI3 - Comparativa con otros clasificadores*

Como se evidencia en la Tabla VI, el rendimiento de los clasificadores supervisados [3] revela varias consideraciones importantes. En primer lugar, se observa que los clasificadores supervisados como SVM, CNN y BERT, muestran un rendimiento generalmente superior al modelo GPT-3.5 Turbo en términos de precisión, sensibilidad y F1-score. Esto sugiere que los enfoques supervisados pueden beneficiarse de un conjunto de datos etiquetados más extenso y específico para la tarea de clasificación en comparación con el modelo de *zero-shot learning* utilizado por GPT-3.5. El resultado va en línea de los estudios que muestran que los modelos GPT sin entrenamiento específico ofrecen buenos rendimientos, pero siempre por debajo de otros algoritmos preparados específicamente para el problema a tratar [20]

Es crucial considerar la forma de entrenamiento de los clasificadores supervisados y la posibilidad de *overfitting*. Los modelos supervisados, al ser entrenados con datos etiquetados, pueden ajustarse demasiado a los detalles específicos del conjunto de entrenamiento. Dado que los conjuntos de datos son obtenidos de un número de fuentes acotado, el rendimiento de estos clasificadores en entornos de producción puede verse mermado al clasificar documentos de otras fuentes. Por otro lado, el modelo GPT-3.5 se basa en un enfoque *zero-shot*, lo que significa que carece de entrenamiento previo sobre los documentos específicos a clasificar. Esto puede resultar en un funcionamiento más generalizable, manteniendo en entornos reales unos valores de rendimiento similares a los obtenidos sobre el conjunto de datos.

Tabla VI
COMPARATIVA DEL RENDIMIENTO (MEDIA PONDERADA) CON CLASIFICADORES SUPERVISADOS

| Modelo | Precisión | Sensibilidad | F1 |
|---|---|---|---|
| SVM | 91.59% | 91.17% | 91.19% |
| CNN | 88.08% | 87.30% | 87.23% |
| BERT | 92.51% | 92.50% | 92.49% |
| **GPT-3.5** | **83.4%** | **79.7%** | **80.5%** |

Al examinar el rendimiento por categoría, es posible identificar similitudes y diferencias entre los clasificadores. Aunque SVM, CNN y BERT muestran un rendimiento consistente en todas las categorías, el modelo GPT-3.5 muestra una variabilidad más pronunciada. Esto sugiere que el modelo GPT-3.5 podría alcanzar un rendimiento cercano al de los modelos supervisados si contase con un entrenamiento específico para la clasificación de ciertas categorías. Este entrenamiento se podría proporcionar a través del *prompt* bajo un enfoque de *few-shot prompting*, como se aplica en [13], o a través de un proceso de *fine-tunning*, que ha demostrado ser efectivo con otros modelos de lenguaje [6].

Ciñéndose a un enfoque *zero-prompt*, el modelo GPT 3.5 Turbo también puede destacar especialmente como clasificador binario en aquellas clases donde discrimina mejor, como drogas o apuestas, poniéndose al nivel de algunos clasificadores semi-supervisados que, con escaso entrenamiento, ofrecen buenas métricas en determinados problemas de clasificación binaria [1], [19]. Para clasificación multi-clase, el modelo GPT-3.5 Turbo presenta una alternativa viable en situaciones donde hay una escasez de datos etiquetados o donde la dificultad técnica de entrenar un modelo supervisado pueda ser un obstáculo. Sin embargo, es importante tener en cuenta las diferencias en el rendimiento y las limitaciones de cada enfoque al considerar su aplicabilidad en diferentes contextos y tareas de clasificación.

## VI. Conclusiones y trabajo futuro

Este estudio ha mostrado el rendimiento de GPT-3.5 a la hora de clasificar el contenido de texto de sitios web de la Dark Web en base al dataset CoDA, bajo un modelo de *zero-shot learning* y un *prompt* básico, consistente en la definición de la tarea a realizar y una descripción breve de cada categoría de clasificación. Bajo estos parámetros, el clasificador alcanza un valor F1 ponderado de 80,5%, mostrando su capacidad para clasificar correctamente una gran cantidad de documentos, especialmente en ciertas categorías. Sin embargo, se detectan también diversas dificultades que le impiden alcanzar un rendimiento equiparable al de los clasificadores entrenados mediante aprendizaje supervisado: la superposición conceptual entre categorías, la aparición de determinada terminología en numerosas páginas de la Dark Web, así como la ausencia de un protocolo claro y completo para realizar la clasificación.

Se ha de tener en cuenta, sin embargo, que el rendimiento de los clasificadores supervisados con mejor desempeño, que alcanzan valores F1 del 94%, parten de un entrenamiento sobre una porción de los datos del propio dataset, lo que implica cierto *overfit* a las características de los documentos del dataset. Bajo el enfoque de *zero-shot learning*, el clasificador carece de entrenamiento previo sobre los documentos a clasificar, y, por tanto, se puede asumir que el rendimiento obtenido será más cercano al rendimiento en entornos reales de producción. Un estudio de rendimiento sobre otros datasets de contenidos de la Dark Web, como DUTA (Darknet Usage Text Addresses) [2], mediante *zero-shot prompting*, podría ayudar a confimar esta hipótesis.

También queda pendiente de estudio la optimización del rendimiento de los modelos GPT. La ingeniería de *prompt* o el

entrenamiento supervisado mediante *fine-tunning* podrían mejorar el desempeño del clasificador acercándolo a los modelos supervisados, pero sin necesidad de realizar un entrenamiento sobre el conjunto de datos completo. Además, en esta investigación se ha medido el rendimiento del modelo GPT-3.5, pero en la actualidad existe un modelo más potente, GPT-4, capaz de resolver problemas complejos con mayor precisión que los modelos anteriores [26]. El uso de este modelo parece idóneo para evaluar el rendimiento del clasificador en un problema significativamente más complejo, como sería proporcionar en el *prompt* información detallada sobre el protocolo a seguir para la clasificación, de forma que pudiera realizarla de forma más cercana a los clasificadores humanos.

## AGRADECIMIENTOS

Este trabajo se lleva a cabo en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation). Ha sido financiado por el Ministerio de Ciencia e Innovación, bajo el "Proyecto para el análisis y recuperación de evidencias criminales asociadas a redes ocultas" (PARCHE), con referencia PID2021-125645OB-I00 de la convocatoria 2021 de "Proyectos de Generación de Conocimiento".

## REFERENCIAS

[1] G. Avarikioti, R. Brunner, A. Kiayias, R. Wattenhofer, y D. Zindros, «Structure and Content of the Visible Darknet». arXiv, 7 de noviembre de 2018. doi: 10.48550/arXiv.1811.01348.

[2] M. W. Al Nabki, E. Fidalgo, E. Alegre, y I. de Paz, «Classifying Illegal Activities on Tor Network Based on Web Textual Contents», en *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, M. Lapata, P. Blunsom, y A. Koller, Eds., Valencia, Spain: Association for Computational Linguistics, abr. 2017, pp. 35-43. Accedido: 9 de noviembre de 2023. [En línea]. Disponible en: https://aclanthology.org/E17-1004

[3] Y. Jin, E. Jang, Y. Lee, S. Shin, y J.-W. Chung, «Shedding New Light on the Language of the Dark Web», en *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, y I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, jul. 2022, pp. 5621-5637. doi: 10.18653/v1/2022.naacl-main.412.

[4] J. Devlin, M.-W. Chang, K. Lee, y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», en *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, y T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019, pp. 4171-4186. doi: 10.18653/v1/N19-1423.

[5] Y. Arslan *et al.*, «A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain», en *Companion Proceedings of the Web Conference 2021*, en WWW '21. New York, NY, USA: Association for Computing Machinery, jun. 2021, pp. 260-268. doi: 10.1145/3442442.3451375.

[6] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, y S. Shin, «DarkBERT: A Language Model for the Dark Side of the Internet». arXiv, 18 de mayo de 2023. doi: 10.48550/arXiv.2305.08596.

[7] B. Settles, «Active Learning Literature Survey», University of Wisconsin-Madison Department of Computer Sciences, Technical Report, 2009. Accedido: 7 de marzo de 2024. [En línea]. Disponible en: https://minds.wisconsin.edu/handle/1793/60660

[8] S. Das Bhattacharjee, W. J. Tolone, y V. S. Paranjape, «Identifying malicious social media contents using multi-view Context-Aware active learning», *Future Generation Computer Systems*, vol. 100, pp. 365-379, nov. 2019, doi: 10.1016/j.future.2019.03.015.

[9] B. Settles, «Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances», en *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, R. Barzilay y M. Johnson, Eds., Edinburgh, Scotland, UK.: Association for Computational Linguistics, jul. 2011, pp. 1467-1478. Accedido: 22 de febrero de 2024. [En línea]. Disponible en: https://aclanthology.org/D11-1136

[10] K. S. Kalyan, «A survey of GPT-3 family large language models including ChatGPT and GPT-4», *Natural Language Processing Journal*, vol. 6, p. 100048, mar. 2024, doi: 10.1016/j.nlp.2023.100048.

[11] K. I. Roumeliotis y N. D. Tselikas, «ChatGPT and Open-AI Models: A Preliminary Review», *Future Internet*, vol. 15, n.º 6, Art. n.º 6, jun. 2023, doi: 10.3390/fi15060192.

[12] M. V. Reiss, «Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark». arXiv, 16 de abril de 2023. doi: 10.48550/arXiv.2304.11085.

[13] L. Loukas, I. Stogiannidis, P. Malakasiotis, y S. Vassos, «Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance». arXiv, 28 de agosto de 2023. doi: 10.48550/arXiv.2308.14634.

[14] L. Li, L. Fan, S. Atreja, y L. Hemphill, «"HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media», 2023, doi: 10.48550/ARXIV.2304.10619.

[15] B. Chen, Z. Zhang, N. Langrené, y S. Zhu, «Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review». arXiv, 27 de octubre de 2023. doi: 10.48550/arXiv.2310.14735.

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

[17] I. D. Buldin y N. S. Ivanov, «Text Classification of Illegal Activities on Onion Sites», en *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, ene. 2020, pp. 245-247. doi: 10.1109/EIConRus49466.2020.9039341.

[18] M. Graczyk y K. Kinningham, «Automatic Product Categorization for Anonymous Marketplaces».

[19] H. Alvari, P. Shakarian, y J. E. K. Snyder, «Semi-supervised learning for detecting human trafficking», *Security Informatics*, vol. 6, n.º 1, p. 1, may 2017, doi: 10.1186/s13388-017-0029-8.

[20] J. Kocoń *et al.*, «ChatGPT: Jack of all trades, master of none», *Information Fusion*, vol. 99, p. 101861, nov. 2023, doi: 10.1016/j.inffus.2023.101861.

[21] «Usage policies». Accedido: 28 de febrero de 2024. [En línea]. Disponible en: https://openai.com/policies/usage-policies

[22] J. Ye *et al.*, «A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models». arXiv, 23 de diciembre de 2023. doi: 10.48550/arXiv.2303.10420.

[23] «OpenAI Platform». Accedido: 29 de febrero de 2024. [En línea]. Disponible en: https://platform.openai.com

[24] L. Henrickson y A. Meroño-Peñuela, «Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT», *AI & Soc*, sep. 2023, doi: 10.1007/s00146-023-01752-8.

[25] B. Zierock y A. Jungblut, *Leveraging Prompts for Improving AI-Powered Customer Service Platforms: A Case Study of Chat GPT and Midjourney*. 2023. doi: 10.13140/RG.2.2.17800.29441.

[26] A. Koubaa, «GPT-4 vs. GPT-3.5: A Concise Showdown», preprint, abr. 2023. doi: 10.36227/techrxiv.22312330.v2.