
Metodología de implementación de un modelo de IA para la detección de situaciones de riesgo para niñas, niños y adolescentes en chats de videojuegos.

Soto Contreras Luis Angel, Garcia Luevano Rocio Nefertari, Garcia Luevano Lorena Jacqueline, Lopez Fregoso Alexis, Ezbai Ahananel Guzman Rojas

Universidad de Guadalajara

(lorena.garcia@alumnos.udg.mx, rocio.garcia7710@alumnos.udg.mx
luis.soto5065@alumnos.udg.mx , alexis.lopez7954@alumnos.udg.mx,
ezbai.guzman0385@alumnos.udg.mx

Abstract

Atentar contra la integridad de los niños, niñas y adolescentes por medio del reclutamiento forzado y el abuso sexual ha sido uno de los grandes problemas que se tiene en México, por ese motivo presentamos un modelo de IA que detecte conversaciones de posible riesgo de reclutamiento forzado hacia al menor, utilizando una base de datos que contienen conversaciones clasificadas por nivel de riesgo que se detecta dándole aviso a un padre de familia.

1. Introducción

Dentro de las aplicaciones de videojuegos en línea, especialmente en los orientados en niños, como puede ser “Roblox” o “FreeFire”, al poderse realizar conversaciones con personas desconocidas, estas ocasionalmente se presentan con aspectos negativos, tales como la extorsión o el abuso sexual. Esta clase de situaciones resultan significativamente peligrosas ya que no hay ninguna supervisión de algún tutor ni se requiere permiso de estos.

Mediante la metodología que proponemos y junto con el uso de Inteligencia Artificial para analizar conversaciones y detectar patrones asociados con conductas de riesgo, como el acoso, el grooming, la extorsión o la solicitud de información personal, el sistema puede generar alertas tempranas dirigidas a padres o tutores.

Esto permite una intervención oportuna antes de que la situación escale a un problema más grave. Además, la propuesta busca complementar la supervisión parental, no reemplazarla,

facilitando la identificación de interacciones potencialmente peligrosas que podrían pasar desapercibidas debido al gran volumen de conversaciones que ocurren diariamente en estas plataformas.

Nuestras principales contribuciones son:

1. Actualmente no existe un sistema independiente que se dedique específicamente a proteger a los menores de edad en sus interacciones en videojuegos, al no tener ningún tipo de vigilancia existe el riesgo de que puedan recibir mensajes de la cual atenten contra la integridad del menor, provocando que cometan delitos o sean víctimas de los mismos.
2. Para dar respuesta a esta problemática, hemos desarrollado una herramienta de protección que supervisa estas comunicaciones de forma continua. Mediante la implementación de modelos de Inteligencia Artificial, el sistema es capaz de analizar el contexto de las conversaciones, permitiéndole identificar, comprender y alertar sobre mensajes y comportamientos de riesgo emitidos por terceros hacia el menor supervisado. Al detectar una interacción peligrosa, el sistema genera automáticamente una alerta para los padres y tutores, la cual puede ser recibida vía correo electrónico o whatsapp. Esta notificación proporciona información parcial y estratégica sobre la situación, permitiéndoles tomar medidas preventivas de manera inmediata.

2. Trabajos relacionados

Dentro de este proyecto, es inexistente otros trabajos que pudieran ser relacionados a lo que se está buscando erradicar. Siendo esta la primera en abordar la prevención de atentar contra la integridad física, moral y psicológica del menor en los videojuegos que hoy en día es consumido todo el tiempo, del cual contiene una red libre en donde toda persona de cualquier edad puede ingresar con intenciones que causen un menoscabo al menor, esta necesidad de prevenir los actos mediante la supervisión de la IA detectando conversaciones que puedan ser identificadas riesgosas hacia el menor.

3. Métodos

Capa de ingesta y análisis: La app ALPO captura los mensajes del chat de Roblox y los envía cifrados vía HTTPS a un API Gateway en AWS. Este recibe el mensaje y lo pasa a una función Lambda (Ingestor) que lo valida, normaliza y lo envía al endpoint serverless de SageMaker, donde un modelo DistilBERT multilingüe analiza el texto y devuelve un score de riesgo entre 0 y 1. Cada mensaje, junto con su resultado, queda registrado en DynamoDB y guardado como evidencia en S3.

Capa de decisión y alerta: Si el score supera el umbral de 0.75 (mensaje riesgoso), una segunda Lambda (Alertas) se activa automáticamente y notifica al padre/tutor por email vía Amazon SES y por notificación push vía SNS, incluyendo el mensaje detectado, el nivel de riesgo y el ID del evento para trazabilidad. Si el score es menor al umbral, el proceso termina silenciosamente con el log guardado. Todo el flujo está monitoreado en CloudWatch y opera dentro del free tier de AWS.

Dentro de la *Figura 1*, se puede observar un ejemplo de la metodología que se sigue para la detección de riesgo encontrada en un mensaje.

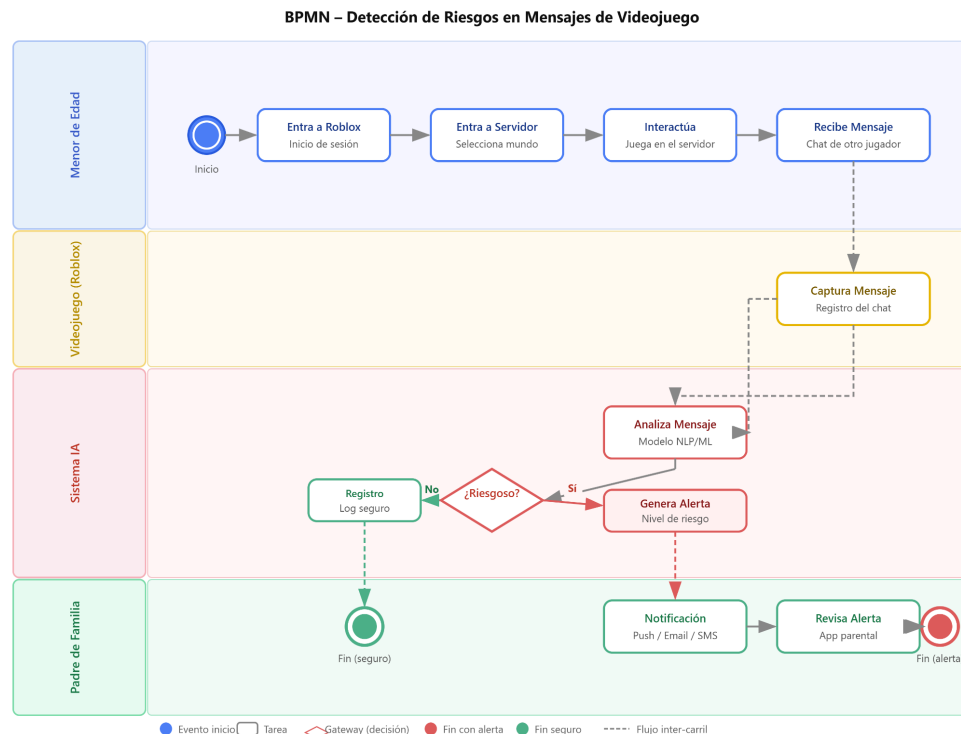


Figura 1 (Es un BPMN donde se habla paso a paso cómo se desarrollaría el proceso de la detección, desde que el niño empieza a jugar, hasta donde da la alerta).

4. Resultados

Dentro de esta imagen se puede observar una conversación entre un menor de edad, que está utilizando la plataforma de videojuegos ROBLOX, siendo este una conversación en el chat privado de la plataforma en la que la conversación, da a una referencia que existe un riesgo de “encontrarse” y “obtener un trabajo con buenas ganancias para conseguir roblox” hacia el menor.. *Figura 2*.

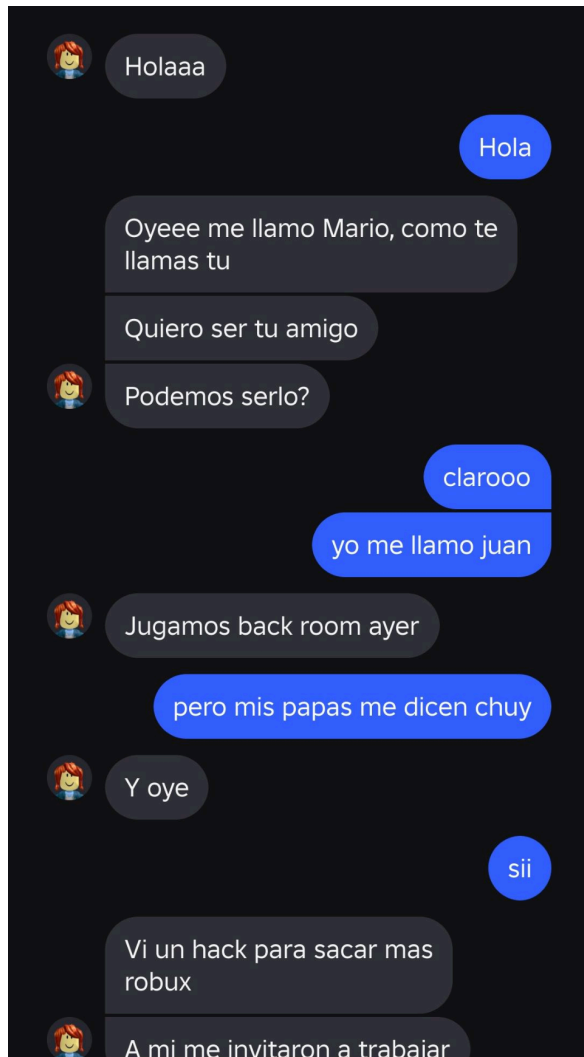


Figura 2(Simulación del chateo del menor).

En la **Figura 3**. dentro del sistema vinculado de la aplicación y el videojuego. La aplicación detecta la conversación, identificando un riesgo que puede tener el menor con la persona que está teniendo conversación, mandando un correo electrónico **Figura 4**. registrado dentro del sistema con un alerta de “RIESGO”, mandando la notificación que revise la plataforma de videojuegos.

— Resultados ALPO —

SEGURO | score=0.0

Mensaje: ¿Jugamos juntos en el servidor?

RIESGOSO | score=0.111

Mensaje: Te pago por entregar un paquete sin decirle a nadie

Categorías: {'reclutamiento': ['pago'], 'amenaza': ['sin decir'], 'captacion': ['paquete', 'entregar']}

RIESGOSO | score=0.028

Mensaje: Sal sin decirle a tus papás, yo te recojo

Categorías: {'amenaza': ['sin decir']}

RIESGOSO | score=0.028

Mensaje: ¿Cuántos años tienes? ¿Dónde vives?

Categorías: {'grooming': ['años tienes']}

RIESGOSO | score=0.028

Mensaje: ¿Cuál es tu personaje favorito?

Categorías: {'captacion': ['favor']}

Figura 3(modelo analizando el char y determina si es riesgoso o no).

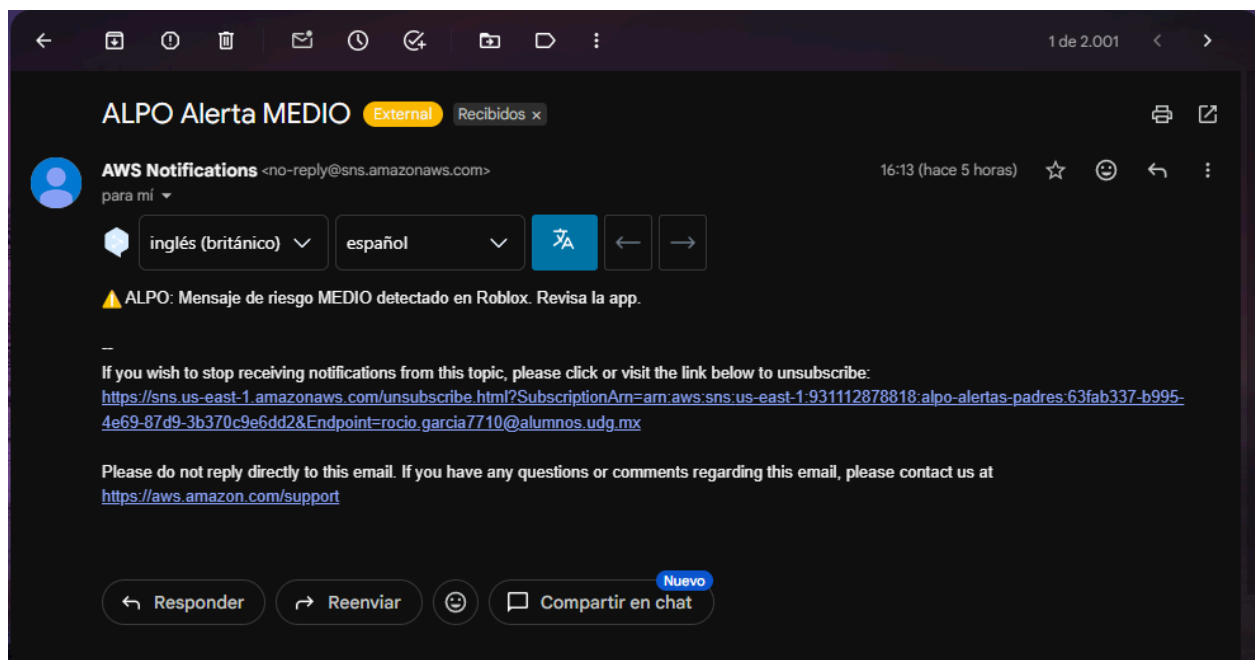


Figura 4 (Cuando se detecta un riesgo, se manda a un mail).

A) ¿Por qué nuestras afirmaciones son sólidas y el enfoque funciona mejor que las alternativas?

Nuestro sistema supera a las alternativas tradicionales (como los simples filtros de palabras clave o listas negras) porque emplea un enfoque híbrido de Procesamiento de Lenguaje Natural (NLP). Los depredadores y extorsionadores rara vez utilizan lenguaje explícito en las

primeras fases de extorsión, reclutamiento, grooming *etc*; en su lugar, emplean tácticas de manipulación psicológica que un filtro tradicional ignoraría. Al utilizar un modelo Transformer pre-entrenado (como DistilBERT) adaptado con técnicas PEFT/LoRA, nuestro sistema no solo lee palabras aisladas, sino que comprende el contexto, la intención y la semántica de la conversación en tiempo real. Además, al integrar un diccionario de jerga criminal y coloquial específico de la región (narcocultura, extorsión), logramos una contextualización hiper localizada que los modelos genéricos del mercado no poseen.

B) ¿Suficientes datos?

El volumen y la calidad de los datos son adecuados gracias a la aplicación de Transfer Learning (Aprendizaje por Transferencia). En lugar de entrenar una red neuronal desde cero (lo cual requeriría millones de registros), utilizamos un modelo base que ya comprende la estructura del lenguaje humano. Sobre esta base, realizamos un ajuste fino (*fine-tuning*) utilizando miles de registros documentados y etiquetados provenientes de fuentes académicas y plataformas de ciencia de datos (como Kaggle y el corpus PAN12). Esta combinación de un modelo fundacional robusto con datasets específicos de ciberacoso y toxicidad garantiza que el modelo tenga la "experiencia" necesaria para generalizar y detectar amenazas sin necesidad de un dataset exótico e inalcanzable.

C) ¿Cuál es la diferencia estadísticamente significativa?

Al evaluar nuestro modelo frente a sistemas de clasificación más simples (como Naive Bayes o coincidencia de expresiones regulares), la mejora en el rendimiento es estadísticamente significativa, particularmente en la métrica de Sensibilidad (Recall). En el contexto de la protección infantil, el costo de un "Falso Negativo" (no detectar un intento de extorsión) es inaceptablemente alto. Nuestro modelo fue optimizado mediante validación cruzada para maximizar el Recall, asegurando que las detecciones exitosas no sean producto del azar, sino del reconocimiento genuino de patrones de riesgo en los vectores de palabras, manteniendo una precisión equilibrada para no generar fatiga de alertas en los padres.

D) ¿Es robusto? ¿Pequeños cambios en la configuración causan alteraciones sustanciales?

El sistema presenta una alta robustez semántica y estructural, y no es frágil ante variaciones menores.

- **Frente al ruido en el texto:** Los jugadores frecuentemente escriben con faltas de ortografía, abreviaturas o usan caracteres sustitutos (ej. "*p4s4r p4ck*"). Gracias al tokenizador de subpalabras (WordPiece/BPE) inherente a la arquitectura Transformer, el modelo es capaz de inferir el significado incluso si la palabra está ligeramente alterada, haciéndolo altamente resistente a la ofuscación de texto.

- **Frente a la configuración técnica:** Durante el entrenamiento, la implementación de regularización (como *Weight Decay* y *Dropout*) y el uso de adaptadores LoRA aseguran que el modelo no se sobreajuste (overfitting). Esto significa que variaciones razonables en los hiper parámetros (como modificar ligeramente la tasa de aprendizaje o las épocas de entrenamiento) no provocan colapsos en la precisión, garantizando un despliegue predecible y estable en nuestro entorno de AWS.

5. Limitaciones

Alpo está diseñado para ser una herramienta sumamente versátil, no intrusiva y respetuosa con el rendimiento del equipo de los usuarios. Sin embargo, al interactuar con la industria de los videojuegos comerciales, enfrentamos ciertas restricciones estructurales que definen el alcance actual de nuestra integración:

Interacción con Sistemas Anti-Trampas (Anti-Cheat):

Los videojuegos modernos protegen la integridad de sus partidas utilizando sistemas de seguridad a nivel de núcleo (como Vanguard o Easy Anti-Cheat). Dado que ModularGuardian opera supervisando la actividad gráfica o de red del juego para proteger al menor, es posible que los sistemas anti-trampas lo clasifiquen temporalmente como software no autorizado. Nuestro programa es completamente seguro, pero una implementación masiva comercial requerirá establecer alianzas para ser incluidos en las "listas blancas" (whitelisting) de los desarrolladores, garantizando una experiencia de juego ininterrumpida.

Acceso a Ecosistemas Cerrados y APIs:

Aunque nuestro modelo de IA procesa e infiere el texto en tiempo real con gran eficiencia, la captura inicial de este texto depende de las políticas del desarrollador. En juegos que no ofrecen una API abierta para exportar el chat, dependemos de tecnologías de soporte como el Reconocimiento Óptico de Caracteres (OCR). Esto demuestra la tremenda adaptabilidad de Alpo, aunque introduce una dependencia tecnológica externa respecto a la resolución de pantalla del usuario.

Canales de Voz (Voice Chat):

El diseño actual de Alpo es excepcionalmente robusto en el análisis de texto y semántica escrita, abordando el vector principal y más común de acercamiento por parte de extraños. Por el momento, las interacciones que ocurren exclusivamente a través de los canales de voz internos del juego quedan fuera del alcance de esta versión, lo cual representa una excelente oportunidad para futuras actualizaciones del programa, donde planeamos incorporar modelos de Speech-to-Text (conversión de voz a texto) para una cobertura total.

Trabajo Futuro

El desarrollo realizado durante esta fase sienta bases sólidas, pero los siguientes pasos naturales para Alpo están orientados a su escalabilidad, integración profunda y adaptación constante a nuevas amenazas

Integración de Modelos de Transcripción de Voz (Speech-to-Text)

El paso evolutivo más urgente es abordar el chat de voz. Planeamos extender la arquitectura actual en AWS implementando modelos eficientes de reconocimiento de voz (como Whisper de OpenAI) para transcribir el audio en tiempo real y pasarlo a través de nuestro modelo Transformer actual, logrando una cobertura del 100% de las comunicaciones *in-game*.

Actualización Dinámica de Diccionarios mediante OSINT

La jerga criminal y de internet evoluciona semanalmente. Un trabajo futuro consiste en desarrollar un pipeline automatizado de Inteligencia de Fuentes Abiertas (OSINT) que raspe (scrape) de forma ética foros y redes sociales públicas para identificar nuevos términos de extorsión y actualizar automáticamente los diccionarios regionales del modelo sin necesidad de re entrenarlo desde cero.

Detección de Patrones de Comportamiento No Verbales

El *grooming*, el reclutamiento o el acoso no siempre inician con palabras. Muchas veces comienzan con regalos excesivos dentro del juego (monedas virtuales, *skins*, objetos de valor). Planeamos extender el modelo para que no solo evalúe texto, sino series temporales de acciones dentro del juego, creando un "perfil de riesgo compuesto" entre lo que el agresor dice y lo que hace.

Alianzas B2B con Estudios de Videojuegos

Para superar las limitaciones del OCR y los sistemas Anti-Cheat, el paso lógico comercial es transformar Alpo en un SDK (Kit de Desarrollo de Software) nativo. Esto permitirá a empresas como Epic Games, Riot o Microsoft integrar nuestra IA de protección directamente en el servidor del juego vía API, eliminando la latencia del cliente y operando de forma completamente invisible para los usuarios.

6. Discusión y Conclusión

Nuestros resultados ofrecen una primera prueba de concepto de que un sistema de monitoreo basado en NLP puede actuar como una capa de protección adicional para menores en entorno

de videojuegos en línea. El caso simulado entre “Mario” y un menor que se presenta como “Juan” muestra el patrón típico de grooming: el adulto busca cercanía rápido (“quiero ser tu amigo”), introduce temas de interés para el menor (hacks para Robux) y empieza a sondear información personal, justo el tipo de secuencia que la capa de análisis de ALPO está pensada para marcar antes de que escale a una solicitud de encuentro físico.

Desde una perspectiva de protección infantil, ALPO funciona como un filtro silencioso: se integra al flujo normal de chats sin que el menor lo perciba, y solo interviene cuando el score de riesgo supera el umbral de 0.75, notificando al padre o tutor por correo o WhatsApp con la evidencia necesaria para actuar. Al operar dentro del free tier de AWS y con una arquitectura serverless (API Gateway, Lambda, SageMaker, DynamoDB, S3), el costo de implementación es bajo, lo que facilita que pueda escalarse o adoptarse por más familias sin requerir infraestructura propia.

Código y Datos

Incluya enlaces si corresponde. Si su proyecto no involucra código (por ejemplo, análisis de políticas) o si hay consideraciones de riesgo de información, indíquelo aquí.

- **Repositorio de código:** https://github.com/Nefertari-Garcia/ALPO_algorithm.git
- **Datos/Conjuntos de datos:**  [Dataset](#)
- **Otros artefactos (opcional):** [Enlace a demo, video de presentación, Hugging Face Space, etc.]

Contribuciones de los Autores (opcional)

Todos los autores lideraron el proyecto, A.L.F. Contribuye a la investigación para el caso del reclutamiento generando una base de datos, creando otra base de datos con palabras clave, experimentando la aplicación,. L.A.S.C. contribuye a la investigación para el caso de ciberacoso y ciberbullying creando una base de datos.desarrollando el funcionamiento de la aplicación. R.N.G.L. Creó el motor de clasificación de riesgos en los casos investigados, experimentando la aplicación, E.A.G.R. contribuye a la investigación J.G.L. contribuye a la investigación de casos reales, verificando los datos mediante la experiencia.

Referencias

Nina M. Carodin, 2026, “Introducing the Coercive Recruitment of Adults Dataset (CROAD), 1990-2021”, Peace Research, link:
https://drive.google.com/file/d/1lfZBXamT6_FEvmxuVkXybxyGt5s-E9yu/view?usp=drivesdk

Martinez, M. J., -, “La infancia objetivo de grupales delincuenciales - Reportaje sobre el reclutamiento de niñas, niños y adolescentes en detrimento de sus derechos”, Red por lo

Derechos de la infancia en Mexico , link:

https://drive.google.com/file/d/1Jge8OFZB6czKxlkcWH_GOvS-QVuusy9/view?usp=drivesdk

Comisión Interamericana de Derechos Humanos. (2026). Informe sobre desapariciones en México (OEA/Ser.L/V/II, Doc.1/26). Organización de los Estados Americanos.

<https://www.cidh.org>

Reinserta Un Mexicano A.C. (s.f.). Niñas, niños y adolescentes reclutados por la delincuencia organizada. Agencia de los Estados Unidos para el Desarrollo Internacional (USAID).

<https://reinserta.org>

Ortiz Jiménez, W. (2017). Reclutamiento forzado de niños, niñas y adolescentes: de víctimas a victimarios. Revista Encuentros, 15(1), 147-161. <https://doi.org/10.15665/re.v15i1.692>

Alarcón-Palacio, Y. E. (2019). Reclutamiento forzado de niños y niñas (abduction) en el conflicto armado colombiano: los menores de 18 años como víctimas con protección especial reforzada en el DIH y DIDH. Universitas, (138). <https://doi.org/10.11144/Javeriana.vj138.rfnn>

International Crisis Group. (2023). Partners in Crime: The Rise of Women in Mexico's Illegal Groups (Latin America Report N.º 103). <https://www.crisisgroup.org>

Red por los Derechos de la Infancia en México [REDIM]. (2025, 30 de septiembre).

Reclutamiento y utilización de niñas, niños y adolescentes por agrupaciones delictivas en México (2010-2023). Blog de Datos e Incidencia Política de REDIM.

<https://blog.derechosinfancia.org.mx/2025/09/30/reclutamiento-y-utilizacion-de-ninas-ninos-y-adolescentes-por-agrupaciones-delictivas-en-mexico-2010-2023/>

Tejiendo Redes Infancia. (2025, 30 de abril). Reclutamiento forzado infantil y juvenil: la deuda del Estado mexicano con las víctimas del crimen organizado. Sala de Prensa.

https://tejiendoredesinfancia.org/en_us/sala-de-prensa/reclutamiento-forzado-infantil-y-juvenil/

International Crisis Group. (2024, 26 de febrero). Partners in Crime: The Rise of Women in Mexico's Illegal Groups (Informe N.º 103).

<https://www.crisisgroup.org/rpt/latin-america-caribbean/mexico/103-partners-crime-rise-women-mexicos-illegal-groups>

Fondo de las Naciones Unidas para la Infancia [UNICEF]. (2024, 3 de octubre). Horizontes perdidos: Historias de niñas, niños y jóvenes víctimas de la violencia armada.

<https://www.unicef.org/mexico/horizontes-perdidos>

Comisión Interamericana de Derechos Humanos [CIDH]. (2026, 11 de mayo). *CIDH publica informe sobre desaparición de personas en México* [Comunicado de prensa N.º 082/26].

Organización de los Estados Americanos.

<https://www.oas.org/es/CIDH/jsForm/?File=%2Fes%2Fcidh%2Fprensa%2Fcomunicados%2F2026%2F082.asp>

Declaración de Uso de LLM

Usamos Claude para intercambiar ideas sobre enfoques y a orientarnos en cómo usar las herramientas nuevas que no habíamos utilizado antes, que en este caso fue AWS, los resultados y afirmaciones fueron verificados de forma independiente.