Prevención del abuso en línea de niñas, niños y adolescentes en América Latina. Evaluación de estrategias de mitigación con Inteligencia Artificial

Nota conceptual del proyecto

1. Introducción

Este documento presenta de manera resumida el trabajo de investigación realizado por el Centro de Internet Seguro - Viguías a través de su Centro de Conocimiento; Incluye las entidades participantes, el objetivo del proyecto, la metodología realizada y, por último, los resultados obtenidos.

2. Organizaciones participantes del proyecto

a. Autores
   i. Universidad de los Andes. Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA).
   ii. Aulas en Paz.

b. Colaboradores
   i. Red Papaz. Centro de Internet Seguro.

3. Nombre del proyecto

Prevention of online abuse of children and adolescents in Latin America and evaluation of mitigation strategies with Artificial Intelligence.[1]

4. Financiamiento

El proyecto fue financiado por el fondo End Violence Against Children (EVAC).[2]

5. Duración: octubre 2021 - abril 2023

6. Problemática

Las violencias digitales contra niñas, niños y adolescentes han tenido un crecimiento acelerado a nivel mundial que demanda atención urgente. En la actualidad, las denuncias de este tipo de abusos son examinadas manualmente por analistas, quienes se exponen constantemente a materiales altamente sensibles que afecta su bienestar y salud mental. A nivel del desarrollo de nuevas tecnologías -por otro lado- a pesar del progreso en las técnicas de procesamiento de lenguaje, hay un vacío en métodos que se especialicen en la detección del abuso infantil con ayuda de dichas técnicas.

7. Objetivo

El proyecto tuvo como propósito desarrollar una herramienta para automatizar el análisis de los reportes de abuso sexual de niñas, niños y adolescentes, buscando así reducir la exposición que tiene el equipo de analistas a este tipo de material.

[1] *Traducción al español: Prevención del abuso en línea de niñas, niños y adolescentes en América Latina. Evaluación de estrategias de mitigación con Inteligencia Artificial*

[2] *Ver más: https://www.end-violence.org/grants/universidad-de-los-andes*

8. Materiales y métodos

En total se empleó la información de 1196 reportes de material de abuso sexual en las categorías de grooming, sextorsión, divulgación de contenido sexual, y ciberacoso. Estos datos corresponden a los reportes recibidos por Te Protejo de Red PaPaz desde enero de 2021 hasta diciembre de 2022. Para asegurar la privacidad y anonimato de los reportes, toda la información referente a los datos personales de cada caso fue eliminada.

En términos generales, para la realización del modelo, se siguieron los siguientes pasos: (i) recopilación de los datos correspondientes a los reportes de Te Protejo; (ii) desarrollo de un sistema de anotación robusto para entrenar el modelo predictivo; (iii) revisión de los datos obtenidos a través de métricas de evaluación; (iv) entrenamiento de los modelos individuales en tres dimensiones: categoría, daño y grado de criminalidad[3]; y (v) aumento de los datos (data augmentation technique) a través de técnicas con palabras al azar.

9. Entregables / valor agregado

Como resultado del proyecto, se diseñó un modelo de lenguaje de gran tamaño (Large Language Model) capaz de analizar los reportes de sextorsión, sexting, grooming y ciberacoso sexual. Se diseñó además un nuevo sistema de anotación para los datos obtenidos, que permiten hacer un análisis profundo de los casos. La herramienta desarrollada permite categorizar de manera automática los reportes en tres dimensiones: sujeto, grado de criminalidad, y daño. Como resultado del proyecto se publicó el artículo "Guarding the Guardians: Automated Analysis of Online Child Sexual Abuse"[4], en donde se encuentra la información detallada del proceso de construcción del modelo que aquí se describe brevemente.

A lo largo del proceso, se destacó la importancia de la articulación multidisciplinar en este proyecto para la prevención de explotación sexual de niñas, niños y adolescentes en línea. Esta investigación contó con la participación de un equipo de investigación de Inteligencia Artificial con énfasis en aprendizaje automático (machine learning), investigadores e investigadoras académicas expertas en las violencias contra niñas, niños y adolescentes, y analistas de la línea de reporte Te Protejo.

En este proyecto Red PaPaz fue un aliado clave, ofreciendo al grupo de investigación los datos que hicieron posible el desarrollo de esta nueva herramienta de automatización para dar respuesta a las violencias que enfrentan niñas, niños y adolescentes. Estos primeros avances en la construcción de esta herramienta, apuntan hacia el mejoramiento del bienestar del equipo de la línea de reporte de Te Protejo, el refinamiento del análisis de los casos que se reciben en la línea y a una mayor eficiencia en la gestión y priorización de los mismos.

10. Limitaciones

Debido al número reducido de datos para facilitar el entrenamiento de la clasificación de ciertos tipos de abusos, el modelo no siempre mostró la efectividad deseada, por

[3] Por categoría se entiende las diferentes dinámicas del abuso sexual en línea en niñas, niños y adolescentes, como sextorsión, grooming, sexting, etc. Por daño, se identificaron los diferentes resultados de la explotación sexual, como la generación de imágenes autoproducidas, así como la distribución o almacenamiento de material de explotación sexual de niñas, niños y adolescentes (MESNNA), entre otros. Por grado de criminalidad, se categorizó por diferentes acciones agravantes como la intención de hacer daño, la intención de realizar divulgación pública de imágenes íntimas, entre otros.

[4] Para ver el paper, se puede ingresar a: https://cinfonia.uniandes.edu.co/wp-content/uploads/2023/09/2308.03880.pdf?x64374

ejemplo en los casos de morphing[5]. Es importante recordar que este modelo fue un primer intento de automatizar el análisis de los casos de reporte de la línea de reporte Te Protejo. El modelo se encuentra aún en un periodo de prueba, por lo que se esperan futuras iteraciones que incorporen las mejoras necesarias para lograr el desarrollo de una herramienta realmente eficiente en el análisis de los abusos.

11. Consideraciones éticas

Debido a la sensibilidad de la temática del proyecto, los detalles en los datos y métodos se mantienen bajo estricta confidencialidad. La información se compartirá con las demás personas, entidades, fundaciones y demás interesadas, para mejorar las estrategias de intervención de los componentes del Centro de Internet Seguro, así como para seguir ahondando en el entendimiento de la violencia online contra niñas, niños y adolescentes.

12. Referencias

Para poder acceder a la información publicada sobre este proyecto, puede ingresar al link: https://cinfonia.uniandes.edu.co/wp-content/uploads/2023/09/2308.03880.pdf?x64374

[5] *Por morphing se entiende a MESNNA modificado artificialmente en donde se representan rostros reales y reconocibles de niñas, niños o adolescentes superpuestos sobre cuerpos de adultos que realizan actividades sexuales explícitas.*