UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

Análisis De Datos

Proyecto semestral*:*
ChatSentinel

Javier Cadagán Parra
Gabriel Huerta Torres
Diego Oyarzo Navia

Fecha: 06 de Mayo, 2025

Introducción

El grooming online presenta una gran amenaza, donde individuos con malas intenciones buscan establecer relaciones de confianza con menores de edad con fines de explotación. El análisis exploratorio de conversaciones de chat nos servirá para entender las tácticas empleadas y así poder desarrollar herramientas de detección. Este informe presenta un análisis exploratorio del dataset de conversaciones de grooming **(ChatCoder2.0).** Con el objetivo de describir su estructura, detallar el proceso de extracción y limpieza y presentar hallazgos iniciales con gráficos.

Datos

Origen y descripción del Dataset
Fuente: ChatCoder HomePage (Se obtuvo mediante Wayback Machine)
Formato original: Archivos XML.
Número de archivos: [56 archivos XML].
Variables principales:

1. user_id: Identificador del usuario.
2. datetime: Fecha y hora del mensaje.
3. message: Contenido textual del mensaje
4. role: Rol del usuario (PREDATOR / VICTIM).

Extracción y Limpieza

1. **Extracción**
   Se realizó un procesamiento de cada archivo XML usando la librería *xml.etree*.

   def parse_chatlog(*xml*):
   leer contenido XML
   obtener los usuarios de cada 'PREDATOR' y 'VICTIM'
   iterar cada 'POST' :
   extraer usuario, fecha y hora, mensaje
   asignar rol según usuario
   agregar al diccionario
   retornar DataFrame
2. **Limpieza**

Después de la extracción, nos encontramos con el problema de que ciertos caracteres inválidos en el XML rompían el parseo, para solucionarlo se utilizaron expresiones regulares (librería *re*) para reemplazarlos por caracteres válidos.

Luego, se realizó una limpieza en la columna 'datetime', donde los formatos de fecha y hora no eran consistentes entre los archivos. Para estandarizarlos se aplicó una transformación al formato [%Y-%m-%d %H:%M:%S].

Finalmente, se realizó una limpieza básica del contenido de los mensajes de chat, eliminando entradas vacías, mensajes nulos y caracteres innecesarios como espacios excesivos o saltos de línea. También se normaliza el lenguaje de los mensajes, reemplazando el uso de "slangs", abreviaciones y modismos comunes en entornos online (por ejemplo, "u" cambia a "you"), así se facilita el procesamiento del lenguaje.

<u>Análisis Exploratorio</u>

1. Estadísticas descriptivas

a. Roles

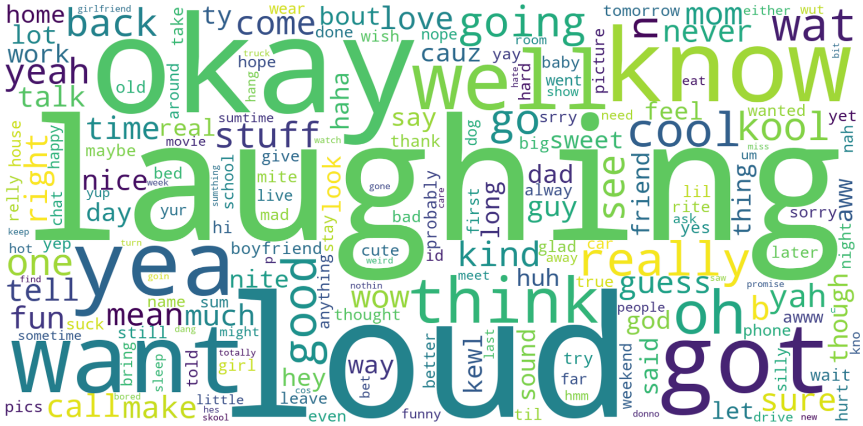
Al DataFrame se le ha agregado una columna que sirve como etiqueta para saber si el mensaje es un groomer o si es una víctima para luego empezar a hacer los análisis correspondientes.

b. Diccionario

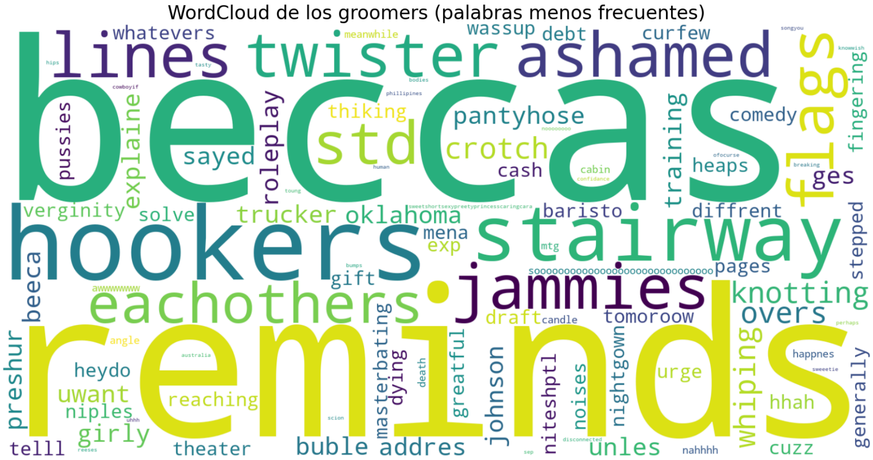
Se ha importado la librería externa ‘contractions’, donde normalizan algunos ‘slang’ de los mensajes en inglés. Esta librería no posee la mayoría de ‘traducciones’ para los slangs de nuestra base de datos, por lo que adicionalmente hemos creado un diccionario de forma manual el cual contiene expresiones regulares para así poder normalizar de manera más completa todos los slangs. Cabe recalcar que el diccionario está en proceso, por lo que algunos slangs no están normalizados.

c. WordCloud (Slangs normalizados con Diccionario)

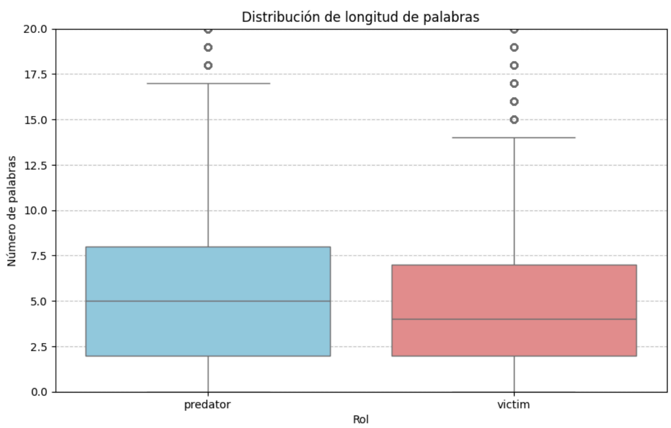
WordCloud de las victimas

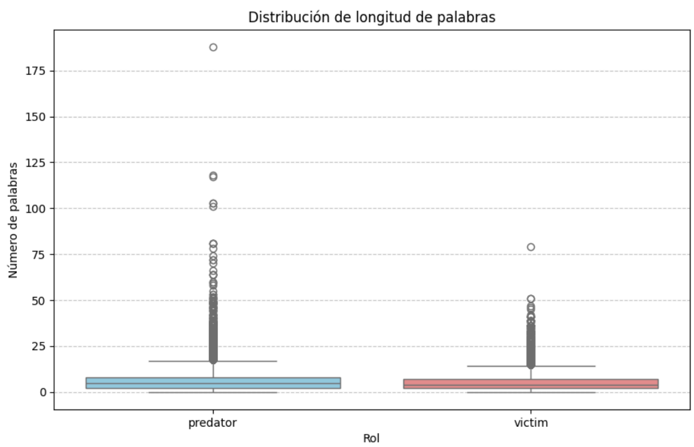


WordCloud de los groomers (palabras menos frecuentes)



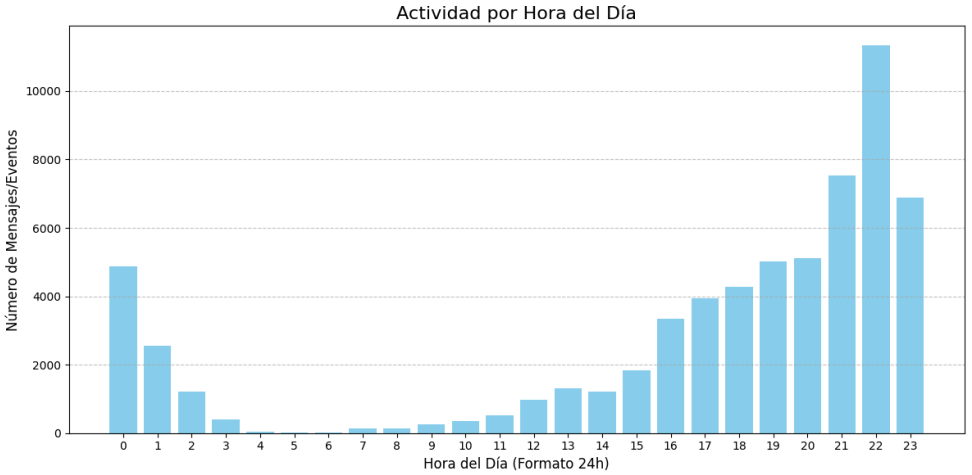
d. Longitud de mensajes





Utilizando la librería de python ‘seaborn’ gráfico, podemos ver que las víctimas suelen usar mensajes más cortos con más frecuencia que los acosadores. Y los acosadores tienen muchos más outliers que las víctimas, de esto último podemos concluir a priori que es porque tratan de manipular a su víctima con mensajes más largos.

e. Tiempo



Este gráfico de barras muestra la cantidad total de mensajes por hora del día (0-23h). El eje X representa la hora del día en formato de 24 horas, y el eje Y muestra el número de mensajes.

El gráfico muestra un patrón claro en la actividad del chat durante el día, podemos ver que la actividad es mínima en horas de la madrugada (3-9h). A partir de las 10h podemos ver que la actividad comienza a incrementar gradualmente, con un aumento más pronunciado a partir de las 15h. El pico máximo de actividad se registra a las 22h con más de 11.000 mensajes y luego se mantiene alto hasta la 1h.

Por lo tanto se sugiere que la mayoría de interacciones en el dataset ocurren durante la noche, esto puede coincidir con los periodos en que los menores de edad tienen más tiempo libre y acceso a dispositivos electrónicos sin supervisión.

Justificación del Uso de los Datos para Abordar el Problema

El análisis del comportamiento lingüístico de ambos perfiles facilita la identificación de patrones comunicativos y tácticas utilizadas por los acosadores. En particular, el análisis de frecuencia léxica mediante *wordclouds* ha permitido visualizar tanto las palabras más usadas como las menos frecuentes por parte de los groomers. Este hallazgo demuestra que existen diferencias significativas en el uso del lenguaje según el rol, lo cual es crucial para el desarrollo de modelos de detección basados en procesamiento de lenguaje natural (NLP). Estas observaciones iniciales respaldan que el dataset no solo es representativo del fenómeno a estudiar, sino también útil desde el punto de vista computacional para diseñar herramientas automáticas de identificación temprana de grooming.