UNIVERSIDAD & SOCIEDAD

Fecha de presentación: diciembre, 2024
Fecha de aceptación:febrero, 2025
Fecha de publicación: marzo, 2025

ANÁLISIS AUTOMATIZADO

DE CONVERSACIONES PEDÓFILAS USANDO MACHINE LEARNING

AUTOMATED ANALYSIS OF PEDOPHILIC CONVERSATIONS USING MACHINE LEARNING

Byron Oviedo-Bayas[1*]
E-mail: boviedo@uteq.edu.ec
ORCID: https://orcid.org/0000-0002-5366-5917
Cristian Zambrano-Vega[1]
E-mail: czambrano@uteq.edu.ec
ORCID: https://orcid.org/0000-0001-8568-8024
Pamela Guevara-Torres[1]
E-mail: pguevarat@uteq.edu.ec
ORCID: https://orcid.org/0000-0001-7863-8678
[1]Universidad Técnica Estatal de Quevedo. Ecuador.
*Autor de correspondencia

Cita sugerida (APA, séptima edición):

Oviedo-Bayas, B., Zambrano-Vega, C. & Guevara-Torres, P. (2025). Análisis automatizado de conversaciones pedófilas usando Machine Learning. ***Universidad y Sociedad, 17***(2), e4996.

RESUMEN

El presente proyecto de investigación se basa en implementar una solución informática para detectar mensajes de texto con intenciones pedófilas a través de aplicaciones de mensajería móvil. Se analiza un corpus de este tipo de conversaciones para seleccionar las características más relevantes utilizando técnicas del procesamiento del lenguaje natural y lograr un mejor desempeño en el modelo predictivo con algoritmos de clasificación supervisada de Machine Learning. Se adoptó el algoritmo de Máquinas de Vectores de Soporte como modelo de clasificación del texto, con las pruebas realizadas a este modelo se obtuvieron resultados muy prometedores con una precisión del 80% en la clasificación de estos mensajes, algunos de estos presentaban mucho ruido y errores gramaticales por lo que se veía afectada la capacidad de aprendizaje del modelo, por lo que luego de realizar una etapa de procesamiento del conjunto de datos y los ajustes necesarios se logró optimizar un poco más el modelo y llegar hasta un 84% en cuanto a precisión y exactitud, además del puntaje F1 que indica un desempeño del 86% para el modelo de clasificación construido. Finalmente, el modelo se implementa en un Bot que se agrega a un grupo de Telegram conectando con su API para el análisis de conversaciones de prueba con el fin de observar que tan bien realizaba la clasificación automática de los mensajes que se enviaban en el grupo.

Palabras clave: Acoso sexual en línea, Aprendizaje automático, Clasificación automática, Mensajería móvil.

ABSTRACT

This research project is based on implementing a software solution to detect text messages with pedophilic intentions through mobile messaging applications. A corpus of this type of conversations is analyzed to select the most relevant features using natural language processing techniques and achieve a better performance in the predictive model with supervised Machine Learning classification algorithms. The Support Vector Machine algorithm was adopted as the text classification model, with the tests performed on this model very promising results were obtained with an accuracy of 80% in the classification of these messages, some of these had a lot of noise and grammatical errors so that the learning capacity of the model was affected, Therefore, after a processing stage of the data set and the necessary adjustments, it was possible to optimize the model a little more and reach 84% in terms of precision and accuracy, in addition to the F1 score that indicates a performance of 86% for the classification model built. Finally, the model is implemented in a Bot that is added to a Telegram group by connecting to its API for test conversation analysis to observe how well it performed the automatic classification of messages sent in the group.

Keywords: Online grooming, Machine learning, Automatic classification, Mobile messaging.

INTRODUCCIÓN

La constante evolución tecnológica ha hecho posible que personas en todo el mundo se comuniquen con otras personas de diversas maneras, independientemente de dónde tengan acceso a Internet, lo cual representa un gran beneficio para la sociedad. Sin embargo, como es de esperarse se puede encontrar muchos riesgos mientras se navega por Internet, los ataques a la privacidad de niños y adolescentes se han incrementado significativamente en los últimos años a través de los medios tecnológicos. Uno de estos peligros es el grooming, el cual consiste en hacer "amigos" en línea, donde un adulto fingiendo ser alguien más entabla una relación de amistad con la víctima, con el objetivo de conectar emocionalmente con el menor de edad hasta ganar su confianza, y así poder obtener contenido erótico o involucrarle en actividades sexuales.

En estos términos, se establece como objetivo crear una solución informática que logre detectar automáticamente este tipo de intención maliciosa en mensajes a través de una aplicación de mensajería móvil y prevenir que menores de edad sigan siendo víctimas de este ciberdelito. Para esto, se indaga primero en diversas fuentes confiables, como trabajos realizados previamente, artículos y revistas científicas, bases de datos y libros, para posteriormente con las bases ya definidas, elegir qué tecnologías permiten el desarrollo del programa. Para finalizar, se lo integrará en una aplicación de mensajería móvil recreando algunas conversaciones de prueba con el fin de comprobar el funcionamiento del proyecto.

MATERIALES Y MÉTODOS

Debido a que se va desarrollar un programa que automatice la detección de contenido grooming en una conversación, se tendrá que realizar varias pruebas controladas para el diseño del modelo de clasificación con los algoritmos que ofrece el aprendizaje automático, además del preprocesamiento del dataset.

Preparación del conjunto de datos

Los algoritmos de clasificación supervisada necesitan de un conjunto de datos preparado para poder aprender de ellos y realizar predicciones sobre nuevos datos.

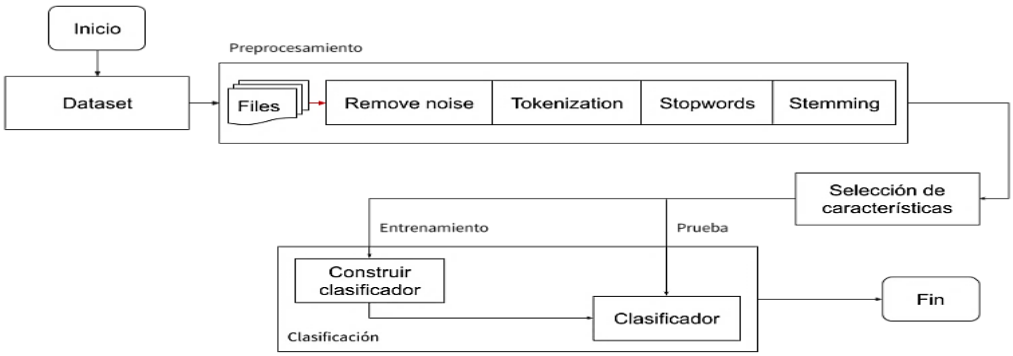
Para el entrenamiento del modelo de clasificación binaria se necesitan de dos clases (1 y 0), por consiguiente, se recolecta conversaciones de tipo grooming y no-grooming.

Las conversaciones de tipo grooming se obtienen del sitio web www.perverted-justice.com, utilizado también por investigaciones anteriores (Beltrán & Ordoñez, 2014; Meyer, 2015; Pendar, 2007; Vogt, 2020; Zambrano et al., 2019;). Mientras que el otro tipo de conversaciones de obtuvo de sitos como www.kagle.com y www.irclog.org, como todos estos datos se encontraban en idioma inglés y parte del propósito es trabajar con datos en español, se recurrió a traducir todo el dataset y eliminar algunos mensajes que carecían de sentido. En total se obtuvo un dataset de 17000 mensajes.

Preprocesamiento

El texto de las conversaciones posee mucho ruido que debe eliminarse antes de determinar las características necesarias para realizar la clasificación, ver figura 1.

Fig 1. Proceso de la investigación.



Fuente: elaboracion propia.

La figura 1 muestra el procedimiento del trabajo que se sigue para construir el modelo de clasificación, distinguiendo como principales:

Files: todos los registros de conversaciones se unieron en uno solo para ambas clases al realizar el preprocesamiento.

Remove noise: se eliminan caracteres especiales y alfanuméricos, signos de puntuación, espacios dobles, abreviaciones, etc.

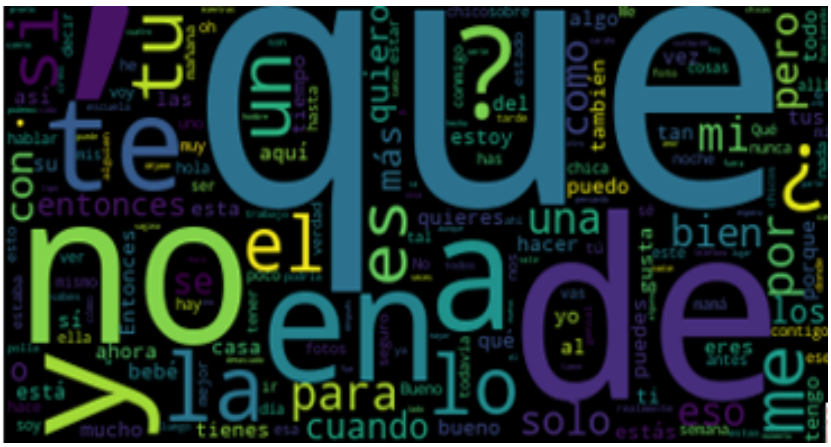
Tokenization: se divide el texto en pequeñas unidades de palabras.

Stopwords: se eliminan las palabras que no aportan mayor información al texto como conectores, pronombres, artículos, etc.

Stemming: se transforma cada palabra a su forma base.

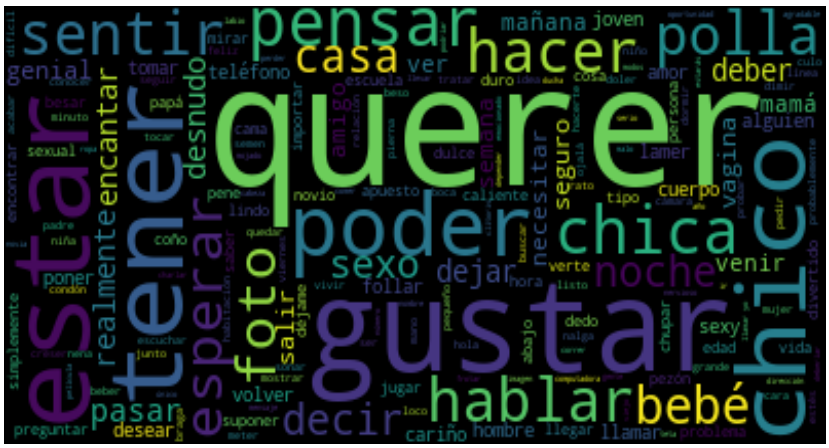
La figura 2 y 3 muestran la diferencia entre un texto procesado y no procesado.

Fig 2. Texto sin procesar.



Fuente: elaboración propia.

Fig 3. Texto procesado.



Fuente: elaboración propia.

Como se observa en la figura 3, únicamente se deja las palabras más frecuentes en todo el dataset de conversaciones de tipo grooming.

Extracción y selección de características

Una vez el texto ha sido preprocesado se transforma en un modelo de espacio vectorial, puesto que los algoritmos de aprendizaje automático son funciones matemáticas necesitan de datos numéricos para realizar los cálculos. Entonces,

se utiliza el algoritmo de vectorización Frecuencia de Términos-Frecuencia Inversa de Documentos o TF-IDF para asignar un valor numérico a cada palabra en cada mensaje del dataset (F1, F2, F3), este valor se calcula de la siguien-

$$TF(t,d) = \frac{frecuencia\ de\ t\ en\ documento\ d}{palabras\ totales\ en\ docmento} \quad (F1)$$

$$IDF(t) = \log\left(\frac{documentos\ totales}{docmentos\ con\ el\ término\ t}\right) + 1 \quad (F2)$$

$$TF\_IDF(t,d) = TF(t,d) * IDF(t) \quad (F3)$$

te manera (Zhao et al., 2018)feature words extraction and topic modeling based on Term Frequency times In-verse Document Frequency (TFIDF:

Donde TF es la ecuación para hallar la frecuencia del término deseado e IDF es la frecuencia inversa del documento, del producto de estos términos se obtiene la medida TF-IDF para inferir la relevancia de cada término en el dataset y así determinar las características más importantes en todo el dataset.

Algoritmo de clasificación

Para problemas de clasificación de textos se utilizan clasificadores lineales por su simplicidad y bajo costo computacional. Uno de estos algoritmos más utilizados para clasificadores es Máquinas de Vectores de Soporte (SVM), puesto que presenta buenos resultados en comparación de otros como se encuentran en las siguientes investigaciones (Beltrán & Ordoñez, 2014; Gunawan et al., 2016; Pendar, 2007; Rivero, 2017; Vogt, 2020).

Aunque estos clasificadores lineales no manejan datos complejos, son suficientes para trabajar con datos de contenido textual como es en este caso. Además, presenta ventajas de rendimiento y eficiencia al momento de entrenar y probar el modelo, esta es una de las razones por la que se optó por utilizar este algoritmo para la clasificación del texto.

RESULTADOS Y DISCUSIÓN

Después de la recolección, la normalización y el procesamiento de los datos, se entrena el modelo de clasificación con el algoritmo SVM y se evalúa su desempeño con nuevos datos. Los resultados de esta evaluación del modelo describen una precisión del 84%, siendo una medida considerablemente buena. La tabla 1 presenta el reporte de clasificación del modelo para ambas clases de etiquetas esperadas.

Tabla 1. Reporte de clasificación del modelo.

	Precisión	Recuerdo	Puntaje F1	Soporte
No groomer	0.84	0.78	0.81	1503
Groomer	0.83	0.88	0.86	1897
Promedio total	0.84	0.84	0.84	3400

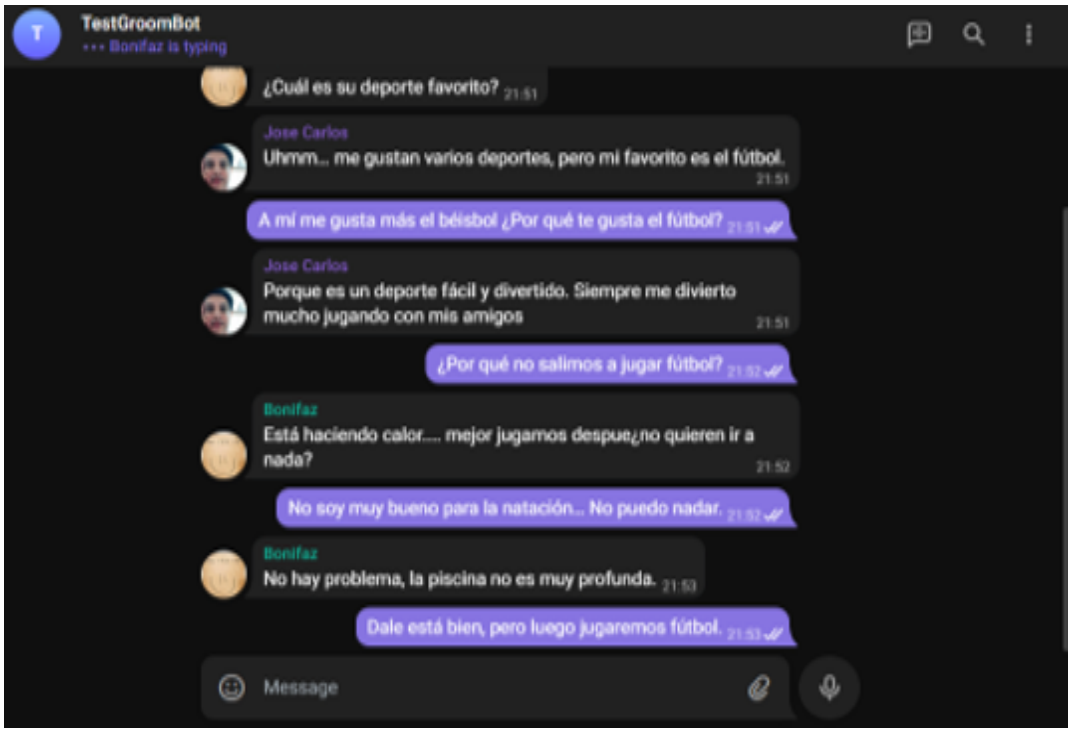
Fuente: elaboración propia.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| = 0.16 \quad (F4)$$

De igual modo, se calcula el error absoluto medio (F4), del modelo el cual es el promedio de la diferencia absoluta entre los valores predichos los reales y, lo ideal es que este valor se lo menor posible para asegurarnos que el modelo tiene un buen ajuste.

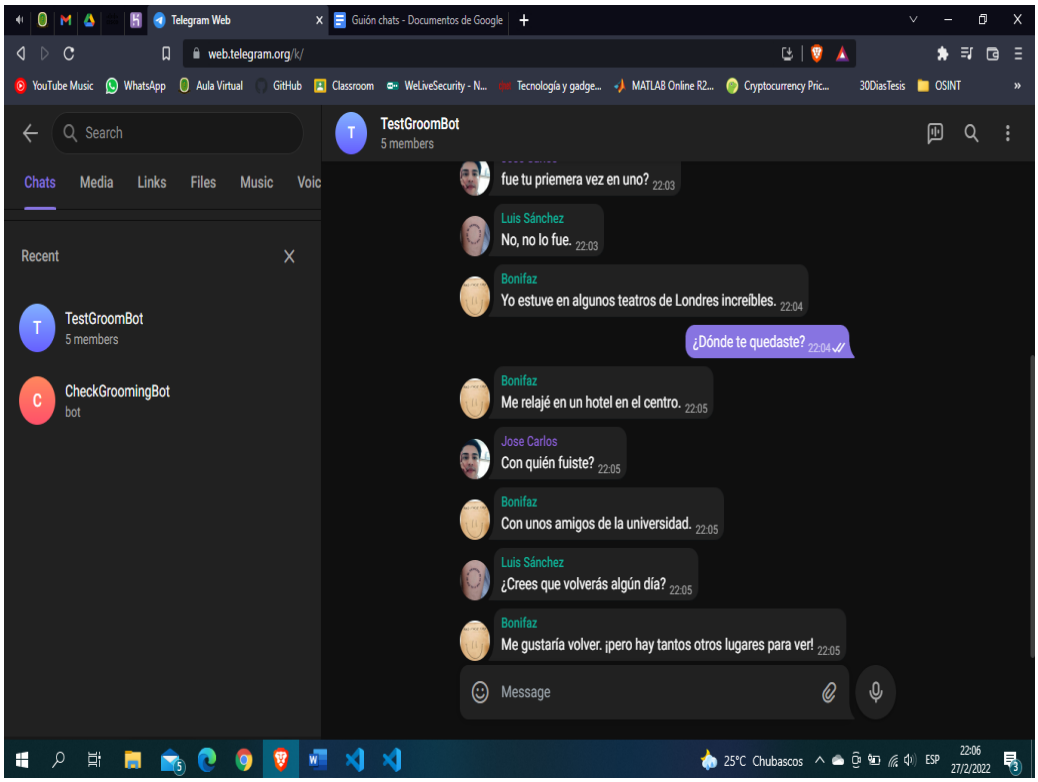
Por último, se simulan 4 escenarios de conversaciones en la aplicación de Telegram dónde se integra el modelo en un Bot, el cual se implementa en un grupo para analizar los mensajes que se envían por ese medio (Figuras 4, 5, 6 y 7).

Fig 4. Chat escenario 1 – Conversación sobre deportes.



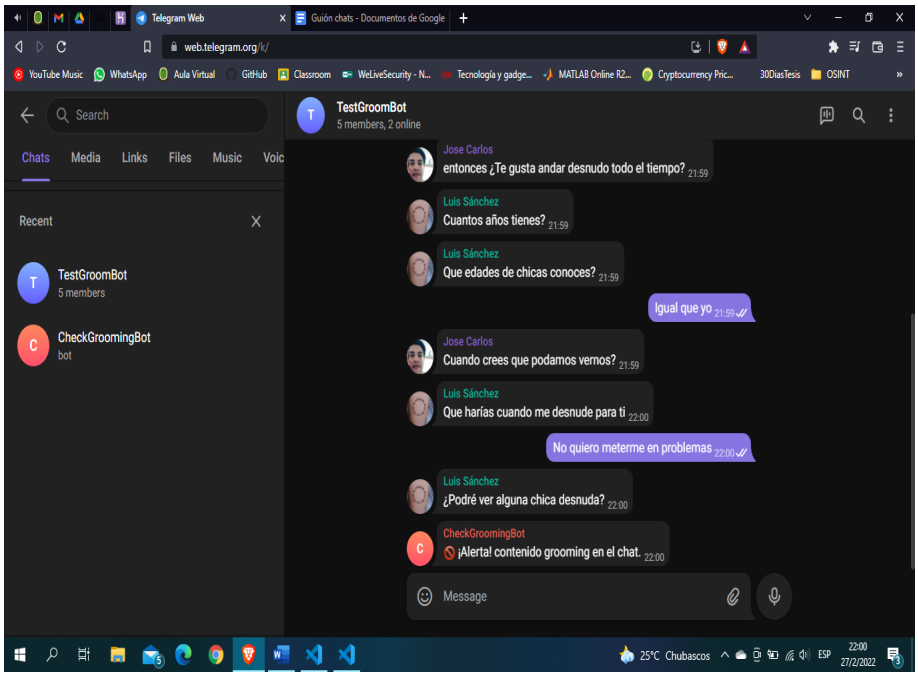
Fuente: elaboración propia.

Fig 5. Chat escenario 2 – Conversación sobre viajes.



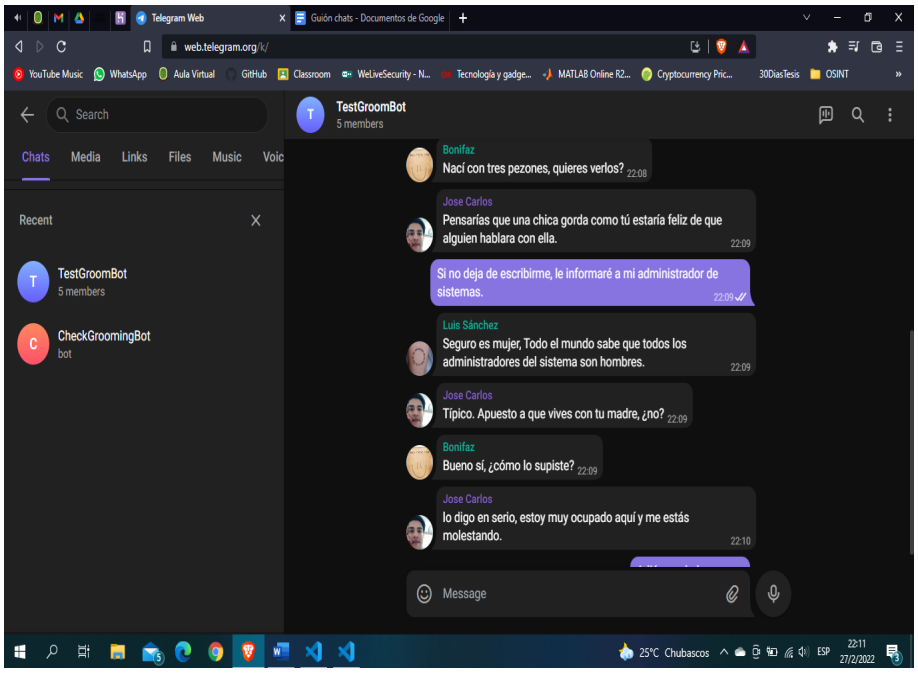
Fuente: elaboración propia.

Fig 6. Chat escenario 3 – Conversación con grooming.



Fuente: elaboración propia.

Fig 7. Chat escenario 4 – Conversación con bullying.



Fuente: elaboración propia.

Cada uno de los mensajes en cada escenario posee un porcentaje de contenido grooming, este porcentaje se promedia con el total de mensajes en la conversación y se obtiene un porcentaje referencial de la cantidad de grooming en el chat, mientras este porcentaje sea menor al 80% no se genera la alerta en la conversación como se expone en la tabla 2.

Tabla 2. Resumen de resultados en cada escenario.

Escenario	% Grooming	Genera alerta
1	49.1	No
2	60.1	No
3	95.9	Si
4	74.4	No

Fuente: elaboración propia.

Con los resultados obtenidos se puede evidenciar que efectivamente con el algoritmo Máquinas de Vectores de Soporte se logran mejores resultados en la clasificación de texto. Acorde las investigaciones anteriores tomadas como referencia se conoce que es posible lograr una precisión mayor al 75% (Meyer, 2015), y ciertamente se supera ese porcentaje en este trabajo. Además, se deja en claro el riesgo que presenta el grooming en el ámbito social a causa del uso no controlado de Internet por parte de jóvenes que llegan a ser víctimas sexuales por este ciberdelito (Machimbarrena et al., 2018)2018.

Otras investigaciones como Oña & Proaño (2020) y Zambrano et al. (2019) obtienen una precisión del 97% y 94% respectivamente, mientras que en un estudio propio anterior se alcanza un 90% de precisión. Cabe resaltar que estos trabajos mencionados utilizan el dataset en idioma inglés y de esta manera se entrena el modelo de clasificación con mejor rendimiento puesto que a diferencia el idioma español es un poco más complejo por ciertas reglas gramaticales entre otras particularidades.

Existen muchas investigaciones sobre este tema del grooming en línea, en la mayoría se lo estudia como un fenómeno psicológico dentro de la criminología como un vector de ataque de ingeniería social siguiendo unos patrones al momento de que uno de estos actores maliciosos intenta atrapar a un menor de edad a través del diálogo hasta ganar su confianza (Lorenzo-Dus et al., 2020). Ahora, conforme los avances tecnológicos en el campo de la inteligencia artificial van madurando, se amplían métodos más potentes para el análisis de textos para que un ordenador identifique de manera automática este proceso de acercamiento sexual por parte de un adulto hacia niños, aunque este procedimiento pueda variar (Kloess et al., 2019).

CONCLUSIONES

Se entrenó un modelo de aprendizaje supervisado a partir de registros de conversaciones en la web que se obtuvieron del idioma inglés, motivo por el cual se recurrió a utilizar servicios de traducción como translate.google.com y deepl.com para garantizar una traducción mucho más confiable y correcta de la gran cantidad de conversaciones que se descargaron, puesto que el objetivo del trabajo es entrenar el modelo con el conjunto de datos en el idioma español. No obstante, una manera más eficaz sería disponer de una base de datos que contenga este tipo de registros de conversaciones disponible.

La aplicación de mensajería Telegram ofrece una API que permitió crear un Bot en el mismo lenguaje Python con el que se construyó el clasificador para el análisis de los mensajes en esta aplicación obteniendo como resultado una precisión del 84% para la detección de mensajes con intención grooming presentes en un chat de la aplicación.

Como trabajo futuro se planea optimizar este modelo de clasificación y evaluar su eficacia en otras aplicaciones de mensajería móvil que se utilicen comúnmente.

REFERENCIAS BIBLIOGRÁFICAS

Beltrán Gómez, A., & Ordoñez Salinas, S. (2014). Sistema inteligente para la detección de diálogos con posibles contenidos pedofílicos∗ Intelligent System for Detecting Dialogues with Possible Pedophilic Contents Système intelligent pour la détection de dialogues avec possibles contenus pédophiles. ***Revista Virtual Universidad Católica del Norte, 42***, 164-181. https://www.redalyc.org/pdf/1942/194230899012.pdf

Gunawan, F. E., Ashianti, L., Candra, S., & Soewito, B. (2016). Detecting online child grooming conversation. (Ponencia). ***11th International Conference on Knowledge, Information and Creativity Support Systems.*** Yogyakarta, Indonesia.

Kloess, J. A., Hamilton-Giachritsis, C. E., & Beech, A. R. (2019). Procesos delictivos de acoso sexual y abuso de menores en línea a través de plataformas de comunicación por Internet. ***Sexual Abuse***, ***31***(1), 73-96. https://doi.org/10.1177/1079063217720927

Lorenzo-Dus, N., Kinzel, A., & Di Cristofaro, M. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. ***Journal of Pragmatics***, ***155***, 15-27. https://doi.org/10.1016/j.pragma.2019.09.010

Machimbarrena, J. M., Calvete, E., Fernández-González, L., Álvarez-Bardón, A., Álvarez-Fernández, L., & González-Cabrera, J. (2018). Internet Risks: An Overview of Victimization in Cyberbullying, Cyber Dating Abuse, Sexting, Online Grooming and Problematic Internet Use. ***International Journal of Environmental Research and Public Health***, ***15***(11). https://doi.org/10.3390/ijerph15112471

Meyer, M. (2015). ***Machine Learning to detect online Grooming***. Universidad de Uppsala. https://uu.diva-portal.org/smash/get/diva2:846981/FULLTEXT01.pdf

Oña Salazar, C. D., & Proaño Chaviznan, J. G. (2020). Implementación de un prototipo de control parental enfocado en la detección inteligente de ataques de grooming. (Trabajo de titulación). Escuela Politécnica Nacional.

Pendar, N. (2007). Toward Spotting the Pedophile Telling victim from predator in text chats. (Ponencia). ***International Conference on Semantic Computing.*** Irvine, CA, USA.

Rivero Tupac, E. (2017). Detección de contenido malicioso mediante técnicas de Machine Learning en las redes sociales. (Trabajo de fin de grado). Universidad de Buenos Aires.

Vogt, M. (2020). Sexual Predator Identification using Machine Learning on Android. https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/studienDiplomArbeiten/finished/2020/expose_vogt.pdf

Zambrano, P., Torres, J., Tello-Oquendo, L., Jacome, R., Benalcazar, M. E., Andrade, R., & Fuertes, W. (2019). Technical Mapping of the Grooming Anatomy Using Machine Learning Paradigms: An Information Security Approach. ***IEEE Access***, ***7***, 142129-142146. https://doi.org/10.1109/ACCESS.2019.2942805

Zhao, G., Liu, Y., Zhang, W., & Wang, Y. (2018). TFIDF based Feature Words Extraction and Topic Modeling for Short Text. (Ponencia). ***2nd International Conference on Management Engineering, Software Engineering and Service Sciences.*** Bangkok, Thailand.