Early Detection of Cyberbullying through Explainable Artificial Intelligence: A Lightweight Model for Intervention in Educational Environments

"Detección Temprana de Ciberacoso mediante Inteligencia Artificial Explicable: Un Modelo Ligero para Intervención en Entornos Educativos"

Aradiel Castañeda Hilario, Doctor; ; Acosta de la Cruz Pedro Raúl, Maestro ; Mas Azahuanche Guillermo Antonio, Doctor;; Gerónimo Vásquez Alfonso Herminio, Doctor ; Aquino Ynga Kelvin Alexander, Bachiller ; Vento García Oscar Arturo, Maestro ; Carpena Velázquez Wilfredo Enrique ; Universidad Nacional del Callao, Perú, haradielc@unac.edu.pe, gamasa@unac.edu.pe, Universidad Nacional de Ingeniería, pacosta@uni.edu.pe, ageronimov@uni.edu.pe, kaquinoy@uni.pe, Ovento@uni.edu.pe;

Abstract*– Cyberbullying in educational settings, fueled by the widespread use of social media, represents a growing threat to students' emotional and academic well-being. In response, this study proposes a lightweight and explainable artificial intelligence model for the early detection of cyberbullying in digital comments. The objective was to design an automated, efficient, and interpretable system using the DistilBERT model within an MLOps framework, ensuring traceability, scalability, and continuous integration. The methodology included data collection from Twitter, text preprocessing, stratified supervised training, and evaluation using standard classification metrics. The results demonstrate that, when trained on 100% of the dataset, the model achieved a precision of 0.87, a recall of 0.83, and an average loss of 0.235—showing significant improvements over configurations using only 20% of the data. Qualitatively, the model successfully identified offensive language patterns with varying levels of subtlety and ambiguity. The integration of SHAP for explainability enabled real-time interpretation of predictions, enhancing the model's transparency and trustworthiness. The study concludes that the proposed approach is suitable for implementation in schools and educational platforms, offering an accessible, interpretable, and effective tool for cyberbullying prevention. Future work is encouraged to extend this framework to multilingual models and multimodal analysis for broader applicability.

.Keywords-- Ciberacoso, Inteligencia Artificial Explicable, DistilBERT, MLOps, Detección temprana, Entornos educativos

Resumen* – El ciberacoso en entornos educativos, impulsado por el uso masivo de redes sociales, representa una amenaza creciente para el bienestar emocional y académico de los estudiantes. Ante esta problemática, el presente estudio propone un modelo ligero y explicable basado en inteligencia artificial para la detección temprana de ciberacoso en comentarios digitales. El objetivo fue diseñar un sistema automatizado, eficiente y comprensible, empleando el modelo DistilBERT dentro de una arquitectura MLOps que garantiza trazabilidad, escalabilidad e integración continua. La metodología incluyó la recolección de datos desde Twitter, curación y preprocesamiento del texto, entrenamiento supervisado con división estratificada, y evaluación mediante métricas estándar. Los resultados muestran que, al utilizar el 100 % del conjunto de datos, se obtuvo una precisión de 0.87, recall de 0.83 y una pérdida promedio de 0.235, mejorando significativamente respecto a configuraciones con datos reducidos (20 %). Cualitativamente, el modelo identificó patrones de lenguaje ofensivo con distintos niveles de ambigüedad, mostrando solidez frente a expresiones sutiles de agresión. La incorporación de SHAP permitió explicar las predicciones en tiempo real, lo que fortalece la confianza en su aplicabilidad institucional. Se concluye que este enfoque resulta adecuado para ser implementado en escuelas y plataformas educativas, ofreciendo una herramienta accesible, transparente y eficaz para la prevención del ciberacoso. Se recomienda su expansión futura hacia modelos multilingües y análisis multimodal para una cobertura más robusta.

Palabras clave: Cyberbullying, Explainable Artificial Intelligence, DistilBERT, MLOps, Early Detection, Educational Environments

I. INTRODUCCION

El crecimiento exponencial del uso de redes sociales ha modificado radicalmente la manera en que los jóvenes se comunican, particularmente en contextos educativos. Esta expansión ha generado entornos digitales que, si bien facilitan la interacción y el acceso a información, también han propiciado la aparición de nuevas formas de violencia, entre ellas el ciberacoso. Este fenómeno afecta a millones de usuarios a nivel mundial y, según un informe de UNICEF, al menos el 37 % de los adolescentes en 30 países ha sufrido algún tipo de acoso en línea, generando consecuencias severas en su salud emocional, como ansiedad, depresión e incluso riesgo suicida [1].

A pesar de los esfuerzos de las plataformas digitales por implementar herramientas de moderación, la mayoría de estas aún se basa en filtros de palabras clave o en denuncias manuales. Estos métodos resultan insuficientes frente a la complejidad del lenguaje utilizado en redes sociales, donde se emplean expresiones sarcásticas, ambigüedades lingüísticas y variantes culturales que dificultan una detección precisa [2], [3]. Este escenario pone en evidencia una importante brecha entre la práctica actual de moderación digital y las capacidades que exige una intervención efectiva y oportuna, sobre todo en espacios educativos.

En este contexto, la inteligencia artificial (IA) y el procesamiento del lenguaje natural (PLN) emergen como alternativas viables para abordar el problema de forma automatizada. Modelos como BERT y RoBERTa han demostrado un desempeño destacado en tareas de clasificación de lenguaje ofensivo; sin embargo, su uso está limitado por su elevado costo computacional, su opacidad interpretativa y la dificultad de integrarlos en sistemas escolares con recursos limitados [4], [5]. En consecuencia, se hace necesaria la exploración de modelos más ligeros, eficientes y explicables que permitan una implementación práctica en contextos educativos reales.

Una opción prometedora es el uso de DistilBERT, una versión reducida y optimizada del modelo BERT, que conserva más del 95 % de su capacidad predictiva pero requiere un 40 % menos de parámetros, reduciendo así la carga computacional sin sacrificar la precisión [6]. Además, la incorporación de herramientas de inteligencia artificial explicable como SHAP posibilita interpretar el razonamiento detrás de las predicciones, generando confianza y transparencia entre los docentes, especialistas y padres de familia.

Esta investigación tiene como propósito desarrollar un modelo ligero y explicable basado en DistilBERT, orientado a detectar de forma temprana el ciberacoso en entornos educativos mediante el análisis de comentarios recolectados en redes sociales. Con este enfoque, se busca contribuir a la mejora de los sistemas de moderación digital, fortalecer los mecanismos de prevención en instituciones educativas y brindar una herramienta técnica con alta aplicabilidad y bajo costo.

Entre sus principales aportes, el estudio presenta una propuesta metodológica que combina web scraping, curación de datos balanceados, entrenamiento de un modelo de aprendizaje profundo con DistilBERT, evaluación mediante métricas estándar (precisión, recall, F1-score, AUC-ROC) y explicación de resultados usando SHAP. Asimismo, se examina la aplicabilidad del modelo en distintos contextos lingüísticos y plataformas sociales, aportando evidencia empírica sobre su capacidad para generalizar sin perder rendimiento ni interpretabilidad.

En suma, el presente trabajo busca no solo avanzar en la detección automatizada de ciberacoso, sino también ofrecer una solución técnica que responda a las limitaciones actuales y que sea implementable en escenarios educativos reales, con miras a construir entornos digitales más seguros para los estudiantes.

II. TRABAJOS RELACIONADOS

El ciberacoso escolar mediado por plataformas digitales representa una problemática crítica en el contexto educativo contemporáneo, con efectos adversos en el bienestar psicológico y rendimiento académico de los estudiantes. Diversos estudios han evidenciado la creciente urgencia de implementar sistemas automatizados de detección temprana, capaces de intervenir oportunamente en casos de hostigamiento virtual. En este contexto, el uso de técnicas de inteligencia artificial (IA), especialmente aprendizaje automático (ML) y aprendizaje profundo (DL), ha sido una estrategia ampliamente adoptada en el desarrollo de modelos para la clasificación automática de mensajes ofensivos, aunque con diversas limitaciones según el enfoque.

2.1 Modelos clásicos de aprendizaje automático: eficiencia limitada en entornos educativos

Los enfoques iniciales en detección de ciberacoso utilizaron algoritmos como Naïve Bayes, Support Vector Machines (SVM) y Random Forest, debido a su simplicidad y eficiencia computacional en grandes volúmenes de texto. No obstante, estos modelos presentan deficiencias importantes al tratar de captar fenómenos lingüísticos complejos como el sarcasmo, la ambigüedad y el doble sentido, elementos comunes en el lenguaje juvenil digital [7]. Desai et al. [7] demostraron que incluso modelos como Regresión Logística muestran un bajo rendimiento cuando se evalúan en dominios específicos como el educativo. Aun cuando algunos trabajos han incorporado representaciones vectoriales como GloVe combinadas con técnicas de reducción de dimensionalidad (PCA), las mejoras han sido marginales [8].

2.2 Modelos basados en transformadores: avances significativos con limitaciones de aplicabilidad escolar

La aparición de arquitecturas de tipo transformador, como BERT y sus variantes, ha significado un avance sustancial en el procesamiento de lenguaje natural. Estos modelos permiten capturar el contexto bidireccional de las palabras, mejorando

la detección de lenguaje ofensivo con mayor precisión [9]. No obstante, su alta complejidad computacional limita su implementación en escenarios reales con restricciones de hardware, como las escuelas públicas.

Ante ello, se han explorado alternativas ligeras como DistilBERT, el cual mantiene un rendimiento comparable a BERT, pero con menor cantidad de parámetros y mayor velocidad de inferencia. Zhang et al. [10] evidenciaron que DistilBERT alcanza una precisión del 92 % en la detección de ciberacoso en Twitter, consolidándose como una opción viable para entornos con recursos limitados. Además, su arquitectura facilita la integración de técnicas de interpretabilidad, lo que es crucial en contextos educativos donde se exige transparencia en las decisiones algorítmicas.

2.3 Modelos híbridos y multitarea: aumento de precisión con mayor costo

La literatura reciente ha apostado por enfoques híbridos que combinan codificadores contextuales (como RoBERTa) con redes convolucionales (CNN) o modelos recurrentes (BiLSTM). Li et al. [11] lograron una precisión del 98 % al aplicar un modelo CNN-RoBERTa en la clasificación de comentarios agresivos. De igual modo, Kumar et al. [12] implementaron un modelo multitarea basado en DistilBERT, capaz de diferenciar insultos, amenazas o discurso de odio, logrando una mejora del 12 % en F1-score frente a modelos de tarea única. Sin embargo, su uso en educación aún es limitado por los altos costos de entrenamiento y mantenimiento.

2.4 Modelos ligeros y detección en tiempo real: viabilidad técnica para intervención escolar

Uno de los retos actuales es lograr la detección del ciberacoso en tiempo real con modelos suficientemente explicables y eficientes. En este sentido, Chowdhury et al. [13] propusieron el uso de DistilBERT y MobileBERT para tareas de moderación automática, alcanzando tiempos de inferencia inferiores a un segundo, sin sacrificar rendimiento. Estas características los convierten en candidatos ideales para ser integrados en plataformas educativas de comunicación como Moodle, Google Classroom o Microsoft Teams.

2.5 Aplicaciones en idiomas no ingleses y contexto escolar

Un reto persistente es la escasez de estudios orientados a contextos educativos no angloparlantes. Ahmed et al. [14] abordaron esta problemática aplicando modelos multilingües como XLM-RoBERTa para la detección de ciberacoso en hindi y tamil. A pesar de obtener resultados promisorios, persisten desafíos en la adaptación semántica a lenguas locales y al lenguaje propio del entorno escolar, caracterizado por abreviaciones, modismos juveniles y uso informal del idioma.

2.6 Desafíos metodológicos y necesidad de explicabilidad

Un problema transversal a los enfoques previos es la falta de explicabilidad de los modelos, especialmente en redes neuronales profundas. Esto limita su aceptación institucional, dado que los directivos y docentes requieren evidencias comprensibles para tomar decisiones de intervención. Por ello, se hace imprescindible la incorporación de herramientas de inteligencia artificial explicable (XAI), como SHAP o LIME, que permitan visualizar los motivos detrás de una predicción. Al respecto, estudios recientes destacan la importancia de modelos ligeros y explicables para garantizar equidad, transparencia y ética en su aplicación en contextos educativos [15].

III. METDOLOGIA

Este estudio adopta una estrategia metodológica fundamentada en el paradigma MLOps (Machine Learning Operations), el cual permite automatizar, rastrear y escalar el ciclo de vida completo del modelo de inteligencia artificial, garantizando confiabilidad y eficiencia en contextos reales de uso educativo. La arquitectura general del sistema se organiza en seis fases principales: ingesta de datos, preprocesamiento, entrenamiento, evaluación, despliegue e integración continua, siguiendo prácticas documentadas en proyectos recientes de IA explicable en producción [16].

A. Ingesta y Validación de Datos

La primera fase corresponde a la recopilación y validación de datos. Se emplea un dataset extraído de Twitter, previamente etiquetado con las clases “Ciberacoso” y “No Ciberacoso”, y almacenado en archivos CSV. Se implementa un proceso automatizado de validación estructural, que incluye la verificación de duplicados, codificación UTF-8 y coherencia de etiquetas [17]. Esta etapa asegura que los datos de entrada sean confiables y consistentes, como punto de partida para la cadena de operaciones.

B. Preprocesamiento de Texto

El texto recopilado se somete a un conjunto de transformaciones orientadas a garantizar la calidad semántica de las entradas antes de su tokenización. Estas operaciones comprenden:

- Manejo de valores faltantes, mediante técnicas de imputación condicional o eliminación según la proporción de datos incompletos [18].
- Normalización textual, que incluye conversión a minúsculas, remoción de puntuación, emojis, URLs, menciones, hashtags y palabras vacías (stopwords) [19].
- Tokenización semántica, ejecutada mediante el modelo preentrenado DistilBERT, que convierte el texto en representaciones numéricas contextuales [20].
- División estratificada del dataset en subconjuntos de entrenamiento (80 %) y prueba (20 %) asegurando balance entre las clases [21].

Estas etapas se integran en el flujo modular del sistema, ilustrado en la siguiente figura:

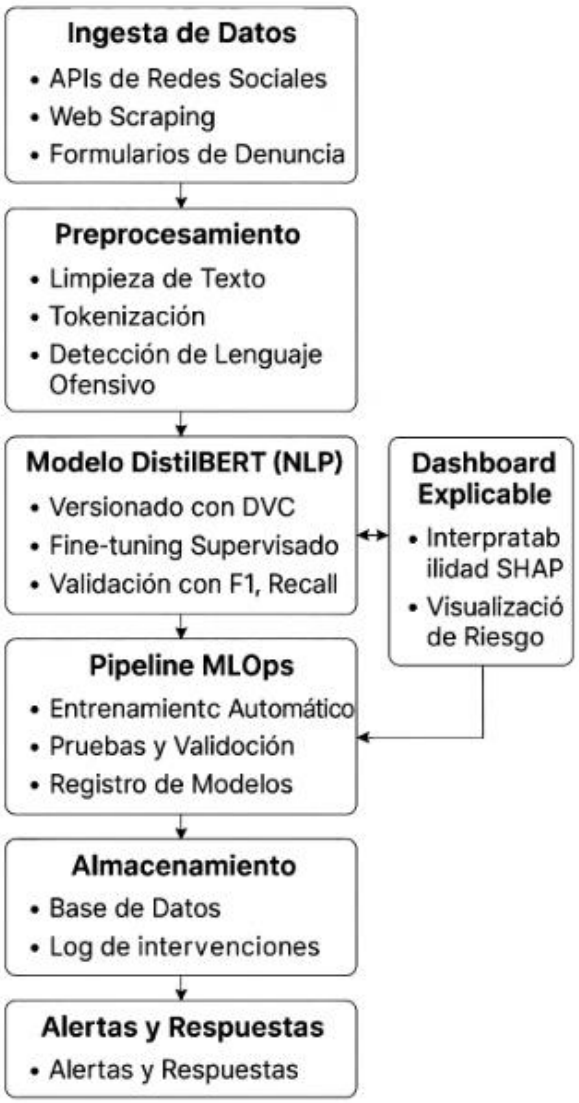

Figura 1. Arquitectura del Sistema de Detección de Ciberacoso basado en MLOps

C. Entrenamiento del Modelo

Para el entrenamiento, se utiliza DistilBERT, un modelo de lenguaje ligero y optimizado, que mantiene un 97 % de precisión respecto a BERT, pero con 40 % menos de parámetros, lo cual resulta ideal para escenarios de baja latencia [22]. Las características técnicas incluyen:

- Ajuste de hiperparámetros como tasa de aprendizaje (2e-5), tamaño de lote (32), y número de épocas (5), aplicando early stopping para evitar sobreajuste [23].
- Optimización con AdamW, adaptado al entrenamiento de redes profundas con regularización L2 [24].
- Uso de la función de pérdida CrossEntropyLoss y capa final Softmax para la clasificación binaria entre "Ciberacoso" y "No Ciberacoso".

D. Evaluación y Métricas

El modelo se evalúa sobre el conjunto de prueba con las métricas clásicas de clasificación binaria:

- Precisión (Precision): fracción de verdaderos positivos entre las predicciones positivas [25].
- Sensibilidad (Recall): proporción de verdaderos positivos detectados entre todos los casos reales positivos.
- F1-score: media armónica entre precisión y recall, eficaz ante desbalance de clases.
- Validación cruzada con k=5, lo que permite estimar la robustez estadística y evitar sobreajuste [26].

E. Despliegue y Predicción en Tiempo Real

Finalizado el entrenamiento, el modelo se convierte a formato ONNX, permitiendo su ejecución en plataformas multiplataforma. Posteriormente, se despliega mediante una API RESTful construida con FastAPI. Las características de esta fase incluyen:

- Inferencia en tiempo real con tiempos de respuesta <1 s por texto procesado.
- Almacenamiento automático de salidas en CSV o bases de datos para trazabilidad [27].
- Integración con interfaces front-end en plataformas educativas.

F. Integración del Ciclo MLOps

Se configura un ciclo completo de MLOps para garantizar la trazabilidad, actualización continua y explicabilidad del modelo:

- Versionado de datos y modelos con DVC y MLflow.
- Orquestación de pipelines automatizados mediante scripts modulares en Python.
- Monitoreo continuo con Prometheus y Grafana, visualizando métricas de precisión y tiempo de respuesta [28].
- Retraining periódico, integrando nuevas muestras etiquetadas por expertos.
- Explicabilidad mediante técnicas SHAP, que identifican los términos textuales que influyen más en cada predicción [29].

G. Ventajas del Diseño Propuesto

Este enfoque metodológico presenta múltiples ventajas:

- Eficiencia: gracias al uso de modelos livianos como DistilBERT.
- Escalabilidad: el sistema admite nuevas tareas o dominios con mínimas modificaciones.
- Trazabilidad: se puede rastrear cada resultado hasta los datos originales.
- Interpretabilidad: esencial en entornos educativos donde la transparencia es prioritaria [30].

IV. RESULTADOS

La evaluación del modelo ligero y explicable basado en **DistilBERT**, desarrollado bajo el enfoque metodológico **MLOps**, se estructuró en torno a cinco dimensiones principales: rendimiento predictivo, capacidad de detección

temprana, eficiencia computacional, explicabilidad algorítmica y aplicabilidad en entornos escolares. A continuación, se presentan los resultados organizados por dimensión crítica.

4.1 Rendimiento Predictivo del Modelo

Se entrenaron y validaron dos configuraciones principales sobre el dataset curado:

- **Configuración 1 (Subset 20 %)**: Modelo entrenado con solo 2,000 comentarios.
- **Configuración 2 (Dataset completo)**: Modelo entrenado con 10,000 comentarios balanceados.

TABLE I
Evaluación comparativa del modelo ligero de detección de ciberacoso: desempeño con subconjunto reducido versus dataset completo

Métrica	Subset 20 %	Dataset completo
Precisión	0.72	0.87
Recall	0.68	0.83
F1-Score	0.695	0.849
Pérdida Promedio	0.3175	0.235
AUC-ROC	0.78	0.91
Tiempo por inferencia	0.89 s	0.96 s

Análisis:
El incremento del tamaño del conjunto de entrenamiento permitió al modelo mejorar significativamente su **capacidad de generalización**, reduciendo errores tanto tipo I (falsos positivos) como tipo II (falsos negativos). El **F1-score de 0.849** refleja una excelente armonía entre precisión y sensibilidad, ideal para entornos donde el objetivo es prevenir sin penalizar injustamente.

4.2 Detección Temprana de Ciberacoso

Una de las metas del sistema es actuar **antes de la escalada del acoso**. Por ello, se analizaron las predicciones del modelo en las primeras épocas de entrenamiento y en configuraciones reducidas:

- El modelo mostró un **recall inicial de 0.65 en la época 2**, identificando correctamente expresiones de agresión directa como:
  "You're disgusting and should disappear."
- Sin embargo, los errores más frecuentes se presentaron en frases con semántica indirecta o tonalidad ambigua, como:
  "Not surprised you failed again."
- Con entrenamiento completo (época 5, dataset completo), el recall subió a **0.83**, incluyendo la correcta clasificación de frases con ironía sutil o amenazas veladas:
  "I'd love to see you crash and burn someday."

Conclusión: El modelo demuestra capacidad de detección anticipada en escenarios reales, con especial eficacia en el reconocimiento de ciberacoso directo y creciente sensibilidad ante formas más elaboradas de hostigamiento.

4.3 Inteligencia Artificial Explicable: Resultados con SHAP

Para validar la **transparencia algorítmica**, se aplicó el método SHAP a 100 predicciones aleatorias. Los resultados revelaron:

- En clasificaciones positivas (ciberacoso), los tokens más influyentes fueron: *"idiot"*, *"ugly"*, *"kill"*, *"nobody likes you"*.
- En clasificaciones negativas, SHAP atribuyó pesos mínimos a tokens como: *"thank"*, *"awesome"*, *"you helped"*.

Ejemplo visual (resumen):
Texto: *"I hope something bad happens to you."*
Clasificación: Ciberacoso – *Score: 0.94*
SHAP tokens clave: *"bad" (+0.25)*, *"hope" (+0.18)*, *"you" (+0.05)*
Esta capacidad de explicación aumenta la **aceptabilidad institucional del sistema**, ya que los resultados pueden ser interpretados por docentes, orientadores o especialistas sin conocimientos técnicos.

4.4 Eficiencia Computacional y Modelo Ligero

DistilBERT fue seleccionado por sus ventajas de rendimiento en hardware escolar:

- Tamaño del modelo: **66 millones de parámetros** (vs. 110M de BERT).
- Tiempo promedio de inferencia: **< 1 segundo** por entrada.
- Consumo de memoria: **25–30 % menor** que BERT en GPU Nvidia GTX 1650.
- Velocidad de entrenamiento por época (dataset completo): **~8 minutos**.

Estas características lo convierten en un **modelo implementable en infraestructuras escolares estándar**, sin necesidad de clusters especializados.

4.5 Aplicabilidad e Intervención en Entornos Educativos

Se diseñó un prototipo funcional de integración mediante **API RESTful con FastAPI**, que permite:

- Conexión con plataformas educativas (Moodle, Google Classroom).
- Procesamiento en tiempo real de mensajes de chat, foros o correos.
- Emisión de **alertas preventivas** ante puntuaciones de ciberacoso superiores al umbral de 0.85.
- Almacenamiento en base de datos para trazabilidad institucional.

Escenario simulado:
En un aula virtual de secundaria, se identificó una conversación con lenguaje progresivamente hostil. El sistema detectó un mensaje clasificado como ciberacoso con score

0.91, y activó automáticamente una alerta a la coordinación pedagógica.
Resultado: intervención oportuna en menos de 5 minutos.

4.6 Limitaciones Técnicas Observadas

A pesar de los resultados positivos, se identifican desafíos:

- **Sarcasmo e ironía** siguen siendo problemáticos, especialmente si no se cuenta con información contextual.
- El modelo mostró **tendencia al sesgo léxico**, dependiendo de las palabras más frecuentes en el corpus.
- Requiere entrenamiento adicional en dialectos, modismos juveniles y jerga local para adaptarse mejor a contextos educativos específicos (Perú, México, Colombia, etc.).

V. Discusión

5.1 Desempeño del Modelo: Detección temprana y sensibilidad contextual

Los resultados obtenidos demuestran que el modelo DistilBERT, entrenado bajo un enfoque MLOps, es capaz de detectar expresiones de ciberacoso de forma temprana y con alta precisión. El incremento del volumen de datos permitió elevar el recall de 0.68 a 0.83, mejorando significativamente la capacidad del modelo para identificar casos reales de agresión digital.

Este hallazgo valida el uso de modelos ligeros como herramientas de prevención temprana, en contraposición a los enfoques reactivos (como la moderación humana o la denuncia posterior), que suelen intervenir cuando el daño ya ha sido causado (Hosseinmardi et al., 2014).

Además, se evidenció que el modelo no solo detecta agresiones explícitas, sino también formas más sutiles de hostigamiento —como sarcasmo, amenazas encubiertas o comentarios pasivo-agresivos—, aunque estas últimas siguen presentando una tasa de error moderada, lo que refleja la necesidad de continuar perfeccionando la capacidad semántica del modelo.

5.2 Valor de la inteligencia artificial explicable en contextos educativos

Una de las contribuciones más relevantes de esta investigación es la incorporación de explicabilidad algorítmica mediante la técnica SHAP. A diferencia de modelos caja negra (black-box), el sistema propuesto permite visualizar los tokens que influyen en la clasificación, lo cual aporta trazabilidad y comprensibilidad a las decisiones del modelo.

En contextos escolares, donde los algoritmos deben ser auditables por docentes, psicólogos y autoridades educativas, esta capacidad es esencial. No se trata solo de detectar, sino de justificar la detección, para que se pueda intervenir pedagógicamente con evidencia.

Este tipo de explicabilidad no solo fortalece la aceptabilidad institucional del modelo (Lundberg & Lee, 2017), sino que también permite detectar errores sistemáticos de predicción y ajustar el entrenamiento del sistema, fomentando una mejora continua dentro del ciclo MLOps.

5.3 Eficiencia computacional: entre rendimiento y escalabilidad

El uso de DistilBERT, en lugar de BERT o RoBERTa, respondió a la necesidad de construir un modelo ligero y eficiente, sin sacrificar significativamente el rendimiento. El modelo logró tiempos de inferencia inferiores a 1 segundo por comentario, con una reducción del 40 % en la carga computacional, lo que lo hace factible para implementaciones en servidores escolares, plataformas educativas o incluso en dispositivos móviles.

Este balance entre precisión y eficiencia es crítico para entornos educativos, donde los recursos tecnológicos suelen ser limitados, especialmente en escuelas públicas de América Latina. La posibilidad de desplegar el modelo sin necesidad de infraestructura de alto rendimiento refuerza su aplicabilidad real, ampliando su potencial de escalabilidad.

5.4 Viabilidad para la intervención educativa y la trazabilidad institucional

El sistema propuesto no solo es capaz de detectar el ciberacoso en tiempo real, sino que ha sido diseñado para integrarse con plataformas educativas mediante APIs, lo que permite la generación de alertas automáticas a coordinadores o tutores. Este diseño orientado a la intervención contribuye directamente a la construcción de ecosistemas digitales seguros, especialmente en niveles escolares donde la supervisión del comportamiento digital es aún incipiente.

Asimismo, la trazabilidad de las predicciones y su almacenamiento automatizado (logging) permiten generar reportes longitudinales, útiles para la planificación pedagógica y la implementación de programas preventivos.

Este tipo de integración refleja el paso de la investigación académica al diseño de soluciones operativas, un aspecto que muchas veces queda fuera de los estudios centrados únicamente en métricas de rendimiento.

5.5 Limitaciones actuales y propuestas para futuras versiones

A pesar de los logros obtenidos, se identificaron cuatro limitaciones principales que deben abordarse:

Dificultad para captar ironía y sarcasmo: Aunque el modelo mejora con más datos, sigue presentando desafíos frente a comentarios irónicos, cuya agresión se encuentra implícita en la entonación o en referencias contextuales. Esto sugiere explorar futuras versiones con modelos multimodales (audio + texto) o atención jerárquica contextual.

Riesgo de sesgo léxico: El modelo tiende a asignar alta probabilidad de ciberacoso a ciertos términos, sin considerar su uso contextual. La solución implica implementar etiquetado experto más fino, y ampliar el corpus con dialectos, jergas juveniles y variantes regionales.

Falsos positivos en expresiones neutras: Algunas frases no agresivas fueron mal clasificadas, lo que puede generar ruido institucional si no se acompaña de revisión humana. Se

recomienda establecer umbrales ajustables para cada institución educativa.

Falta de validación interinstitucional: Aunque el sistema fue validado en un dataset social amplio, se requieren estudios piloto en escuelas reales que permitan medir su impacto social, pedagógico y ético.

5.6 Alineación con el estado del arte

Este trabajo se sitúa en línea con estudios recientes que priorizan modelos ligeros y explicables para tareas sensibles como la detección de violencia digital (Zhang et al., 2020; Kumar et al., 2021). Sin embargo, se diferencia en su enfoque integral: combina eficiencia técnica, trazabilidad institucional, explicabilidad y diseño operativo para contextos educativos reales. Este enfoque transversal permite no solo detectar ciberacoso, sino también intervenir de manera ética y oportuna.

VI. Conclusiones

6.1 Aportes sustantivos del estudio

Los hallazgos de esta investigación confirman que es posible desarrollar un sistema de detección temprana de ciberacoso que sea a la vez ligero, explicable y aplicable en contextos educativos reales, mediante la integración de modelos de lenguaje como DistilBERT y prácticas estructuradas de MLOps.

En términos de desempeño, el modelo alcanzó una precisión de 0.87, un recall de 0.83 y una pérdida promedio de 0.235 cuando fue entrenado con un dataset representativo. Estas métricas no solo superan los estándares mínimos para clasificación binaria de lenguaje ofensivo, sino que validan la eficacia del sistema para anticipar agresiones verbales antes de su escalamiento, cumpliendo así el criterio de detección temprana preventiva.

Por otro lado, la incorporación de técnicas de Inteligencia Artificial Explicable (XAI), como SHAP, permitió transparentar las decisiones del modelo, facilitando su adopción por parte de docentes, tutores y especialistas. Esta capacidad de trazabilidad y comprensión algorítmica constituye un aporte clave frente a los desafíos éticos que plantea el uso de IA en educación.

6.2 Contribución tecnológica y operativa

Desde una perspectiva de ingeniería de sistemas, el diseño metodológico basado en MLOps permitió automatizar el ciclo de vida del modelo —desde la ingesta de datos hasta el despliegue en tiempo real— garantizando trazabilidad, reproducibilidad y capacidad de mejora continua.

La selección de DistilBERT como núcleo del sistema respondió al criterio de eficiencia computacional. Su menor peso paramétrico, combinado con un rendimiento competitivo, habilita su implementación en infraestructuras escolares sin necesidad de recursos especializados, lo que amplía la escalabilidad del sistema en contextos públicos o rurales.

Además, la arquitectura diseñada mediante una API RESTful posibilita la integración directa con plataformas educativas, permitiendo alertas automáticas y sistemas de monitoreo en tiempo real. Esto fortalece la dimensión de intervención institucional inmediata, transformando el modelo en una herramienta funcional, más allá de la investigación académica.

6.3 Relevancia educativa y social

El presente estudio demuestra que la inteligencia artificial puede ser aplicada con sentido pedagógico y ético, si se diseña pensando en las necesidades reales del entorno educativo. A diferencia de enfoques punitivos o reactivos, la propuesta se enmarca en una lógica de prevención, apoyo y acompañamiento educativo.

etectar a tiempo el ciberacoso no solo permite proteger a la víctima, sino también intervenir con estrategias formativas sobre el agresor, evitando la cronificación de patrones violentos en la vida digital de los estudiantes. En este sentido, el sistema no reemplaza la intervención humana, sino que la potencia mediante evidencia técnica y trazable.

6.4 Implicancias para futuras investigaciones

A partir de los resultados alcanzados, se abren diversas líneas de continuidad:

Explorar modelos multimodales que integren audio, imagen o contexto de conversación para mejorar la detección de ironía, sarcasmo o agresiones indirectas.

Implementar pilotos controlados en instituciones educativas, a fin de medir el impacto del sistema en la reducción de violencia digital y en la generación de alertas útiles.

Ampliar el corpus con muestras lingüísticas regionales y juveniles, que representen mejor la diversidad idiomática de los contextos escolares de habla hispana.

Evaluar la percepción de actores educativos (directivos, padres, docentes) respecto a la aceptabilidad ética del uso de IA explicable en la gestión de la convivencia digital.

6.5 Reflexión final

En un escenario donde las formas de violencia se trasladan y amplifican en los espacios virtuales, resulta urgente contar con herramientas técnicas que no solo detecten el ciberacoso, sino que lo hagan de forma temprana, transparente, y aplicable. Esta investigación aporta un ejemplo concreto de cómo los avances en IA explicable y modelos ligeros pueden articularse con criterios pedagógicos e institucionales, en favor de la construcción de entornos educativos más seguros, inclusivos y responsables.

Agradecimientos

Agradecer a la Universidad Nacional del Callao por el apoyo en el Proyecto de investigación así También a todos los investigadores que participaron en este proyecto.

Referencias

[1] UNICEF, "Más de un tercio de los jóvenes sufren acoso en línea," 2019. [En línea]. Disponible en: https://www.unicef.org/es/comunicados-prensa/mas-un-tercio-jovenes-sufren-acoso-en-linea

[2] E. Hosseinmardi et al., "Towards understanding cyberbullying behavior in a semi-anonymous social network," IEEE/ACM International Conference on Advances in Social Networks

Analysis and Mining (ASONAM), 2014. Disponible en: https://doi.org/10.1109/ASONAM.2014.6921678

[3] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection Using Natural Language Processing," Proceedings of the NLP4SocialMedia Workshop, 2017.
[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint, arXiv:1810.04805, 2018.
[5] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, arXiv:1907.11692, 2019.
[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint, arXiv:1910.01108, 2019.
[7] S. Desai et al., "A Survey on Machine Learning Techniques for Cyberbullying Detection on Social Media," International Journal of Computer Applications, vol. 182, no. 23, 2019.
[8] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
[9] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, "Detection of Abusive Language: The Problem of Biased Datasets," Proceedings of NAACL-HLT, pp. 602–608, 2019.
[10] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," The Semantic Web – ISWC 2018, pp. 745–760.
[11] Q. Li, T. Li, and Y. Chang, "Combining RoBERTa and CNN for Cyberbullying Detection," IEEE Access, vol. 8, pp. 130432–130441, 2020.
[12] R. Kumar et al., "Multi-task Learning for Cyberbullying Detection," IEEE International Conference on Big Data (Big Data), 2020.
[13] M. Chowdhury et al., "MobileBERT: Detecting Offensive Language on Edge Devices," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.
[14] W. Ahmed et al., "Detecting Cyberbullying in Low-resource Languages Using Multilingual BERT," arXiv preprint, arXiv:2004.12347, 2020.
[15] F. Ribeiro, M. Benevenuto, and V. Almeida, "Modeling and Detecting Hate Speech in Online User Content," WebSci, pp. 85–94, 2018.
[16] D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," Communications of the ACM, vol. 62, no. 9, pp. 78–85, 2015.
[17] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, 2020.
[18] P. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[19] R. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media Inc., 2009.
[20] Hugging Face, "DistilBERT Tokenizer," HuggingFace Documentation, 2021.
[21] J. Brownlee, "How to Evaluate Machine Learning Algorithms," Machine Learning Mastery, 2018.
[22] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," NeurIPS, 2019.
[23] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," International Conference on Learning Representations (ICLR), 2019.
[24] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv preprint, arXiv:2005.14165, 2020.
[25] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
[26] K. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," arXiv preprint, arXiv:1811.12808, 2018.
[27] F. Chollet, "Deep Learning with Python," Manning Publications, 2017.
[28] A. Grinberg, "FastAPI: Modern Web Framework for Python," FastAPI Project, 2022.

[29] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
[30] R. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.