Comunicar, n. 80, v. XXXIII, 2025 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

Percepción y detección del ciberacoso: comparativa entre la comunidad educativa y las GenAI

Perception and Detection of Cyberbullying: A Comparison between the Educative Community and GenAI

Dr. Ruben Nicolas-Sans*, UNIE Universidad (España) (ruben.nicolas@universidadunie.com) (https://orcid.org/0000-0002-9234-5764)
Rocío Navarro Martínez, UNIE Universidad (España) (rocio.navarro@universidadunie.com) (https://orcid.org/0000-0002-6899-0157)

RESUMEN

El acoso escolar implica comportamientos repetitivos y agresivos para intimidar o dañar a otros, siendo el ciberacoso el tipo que se beneficia de las plataformas digitales para transmitir tales mensajes. Ambos tienen consecuencias graves sobre la salud mental y el rendimiento académico de los estudiantes, lo que requiere estrategias de prevención e intervención, incluyendo la capacitación de los docentes y el uso de herramientas para detectar estos comportamientos en los entornos universitarios. La inteligencia artificial (IA), específicamente la IA generativa, puede detectar automáticamente el lenguaje ofensivo en las plataformas digitales, convirtiéndose en una herramienta eficaz para combatir el ciberacoso. Este estudio analiza la capacidad de los miembros de la comunidad universitaria—estudiantes, docentes y personal administrativo—para percibir y detectar el ciberacoso en los mensajes de redes sociales. La investigación utiliza herramientas de IA generativa para evaluar su efectividad en reconocer patrones de ciberacoso, comparando los resultados con las evaluaciones de expertos. Los resultados indican que los docentes son los más efectivos en identificar el ciberacoso, mientras que los estudiantes muestran mayor indulgencia, lo que resalta la necesidad de intervenciones educativas más específicas. A pesar de sus limitaciones, los modelos de IA generativa demuestran un gran potencial para la detección temprana del ciberacoso. Los hallazgos subrayan la importancia de la formación dentro de las comunidades educativas y sugieren que las herramientas de IA, cuando se integran en programas preventivos, pueden mejorar la intervención temprana y promover entornos.

ABSTRACT

Bullying involves repeated and aggressive behaviors to intimidate or harm others, with cyberbullying being the kind that benefits from digital platforms to direct such messages. They both have serious consequences on students' mental health and academic performance, demanding prevention and intervention strategies, including teacher training and the use of tools to detect these behaviors in university environments. Artificial Intelligence (AI), specifically generative AI, can automatically detect offensive language on digital platforms, resulting in an effective tool combating cyberbullying. This study investigates the capacity of members within a university community—students, faculty, and administrative staff—to perceive and detect cyberbullying in social media messages. The research utilizes generative AI tools to assess their effectiveness in recognizing cyberbullying patterns, comparing their results against expert evaluations. Results indicate that faculty members are most effective in identifying cyberbullying, while students show greater leniency, highlighting the need for targeted educational interventions. The generative AI models, despite limitations, demonstrate potential for early cyberbullying detection. Findings underscore the importance of training within educational communities and suggest that AI tools, when integrated into preventive programs, can enhance early intervention and promote safer digital environments.

PALABRAS CLAVE | KEYWORDS

Redes sociales, cyberbullying, educación superior, estudiantes universitarios, tic, inteligencia artificial
Social Media, Cyberbullying, High Education, University Students, Ict, Artificial Intelligence.

Received: 2024-02-28 | Reviewed: 2025-01-30 | Accepted: 2025-03-04 | Online First: 2025-04-18 | Published: 2025-04-20
DOI: https://doi.org/10.5281/zenodo.15564949 | Pages: 1-12

1. Introducción

1.1. El acoso y el ciberacoso en el ámbito educativo

El acoso se refiere a un comportamiento agresivo repetido destinado a dañar o intimidar a otra persona física, ya sea mental o emocionalmente (Cerezo, 2009). Este comportamiento a menudo se caracteriza por una desigualdad de poderes, donde la persona que participa en el comportamiento de acoso lo perpetra desde un lugar de superioridad respecto a la víctima. El acoso puede adoptar muchas formas, que incluyen: el acoso verbal, la agresión física, la exclusión social, la difusión de rumores y el ciberacoso (Crothers y Levinson, 2004). Particularmente, el ciberacoso se caracteriza por un uso deliberado y repetido de este tipo de dinámicas dentro de plataformas de comunicación electrónica, como por ejemplo aplicaciones de redes sociales, mensajería instantánea, correo electrónico o juegos en línea. En otras palabras, el ciberacoso implica el uso de tecnologías digitales para dirigirse a individuos con mensajes perjudiciales o hirientes, amenazas, rumores u otras formas de comportamiento abusivo (Ventura, Rodríguez-García y Reche, 2018).

Entre los jóvenes, el ciberacoso se manifiesta de diversas maneras, incluyendo la publicación de comentarios despectivos, la compartición de información privada sin consentimiento, la difusión de rumores o mentiras, la creación de perfiles falsos para suplantar o burlarse de otro usuario, o la participación en comportamientos excluyentes en línea (Cortés, De los Río y Pérez, 2019). De forma habitual, el ciberacoso ocurre en el contexto de las relaciones interpersonales, a menudo involucrando a amistades, compañeros de clase o conocidos, pero también puede involucrar a extraños o individuos anónimos. De igual forma, este puede tener serias consecuencias negativas para las víctimas, incluyendo angustia emocional, aislamiento social, problemas académicos y, en algunos casos extremos, daño físico o autolesiones (Schneider et al., 2012).

El ciberacoso posee atributos distintos que lo diferencian del acoso escolar tradicional. Habilitados por las comunicaciones electrónicas, los ciberacosadores pueden operar de manera anónima y alcanzar a una amplia audiencia con sus mensajes. Además, el entorno virtual puede mermar el sentido de responsabilidad e infundir un sentimiento de impunidad de cara a una rendición de cuentas para los perpetradores, en comparación con las interacciones cara a cara (Barlett, 2015; Noblia, Renato y Gershanik, 2022). Estas circunstancias sugieren que las personas que pueden no ser sujetos de acoso escolar tradicional, podrían encontrarse siendo blancos de prácticas on-line de estas características. Las investigaciones que examinan la intersección entre el acoso escolar y el ciberacoso arrojan hallazgos, con estimaciones que indican que entre un tercio y tres cuartas partes de los jóvenes experimenta acoso ya sea en línea o en entornos escolares tradicionales (Schneider et al., 2012). Adicionalmente, diversos estudios ponen de manifiesto que, mientras que en el acoso escolar tradicional los chicos tienen más probabilidades de ser víctimas (Kowalski y Limber, 2007; Wang, Iannotti y Nansel, 2009), en el ciberacoso las diferencias de género no están claras (Hinduja y Patchin, 2008; Tokunaga, 2010; Ybarra y Mitchell, 2004). En otras palabras, el acoso en línea es un comportamiento que no discrimina edades, sexo, raza o posición socioeconómica, es generalizado y se presenta en todos los contextos educativos en mayor o menor intensidad (Reynoso, González y López, 2021).

Los expertos alertan de este fenómeno y de las consecuencias que repercuten en los damnificados. Un estudio publicado en el Scandinavian Journal of Public Health (Landstedt y Persson, 2014), mostró que todos los tipos de acoso se asociaron con síntomas depresivos tanto en niños como en niñas; y todas las formas de acoso aumentaron la probabilidad de problemas psicosomáticos en las niñas. Los expertos confirman la necesidad de una estrategia de prevención e intervención para abordar ambas formas de acoso y su relación con la salud mental y el rendimiento escolar (Schneider et al., 2012). Con respecto al rendimiento escolar y las calificaciones, existen multitud de estudios que relacionan tanto el acoso como el ciberacoso con una reducción significativa del rendimiento escolar (Egeberg et al., 2016; Muzamil y Shah, 2016; Yousef y Bellamy, 2015). Existe un vínculo entre el ciberacoso y el desarrollo problemas de salud mental (Lucas-Molina et al., 2022), así como, una relación de dependencia entre el estrés postraumático causado por el acoso y el consumo de sustancias (Houbre et al., 2006). Los efectos psicológicos negativos provocados por una exposición constante a una situación de acoso aumentan el riesgo de que los jóvenes busquen alivio temporal a través de sustancias, lo que da pie a una serie de consecuencias perjudiciales en su desarrollo y bienestar (Pichel et al., 2022).

En este contexto, el rol de los adultos responsables en el entorno es crucial. El rol de los docentes en instituciones educativas es especialmente relevante en la prevención y detección del ciberacoso ya que las intervenciones de los profesores tienen efectos claros en la adopción de roles relacionados con el acoso por parte de los estudiantes y podrían ayudar a dirigir de manera más efectiva las estrategias de

intervención (Burger, Strohmeier y Kollerová, 2022). La capacidad tanto de profesores como de los padres para identificar con precisión los escenarios tradicionales de acoso y ciberacoso ha sido objeto de estudio (Campbell, Whiteford y Hooijer, 2019) e investigaciones recientes sugieren que los docentes necesitan y desean una mayor capacitación sobre el acoso escolar. Por otro lado, con la aparición y rápido avance de la inteligencia artificial, esta tecnología se presenta como una herramienta prometedora para la detección temprana del ciberacoso, como demuestran diversos estudios en la materia orientados a identificar patrones de ciberacoso en redes sociales (Azeez et al., 2021). La ventaja principal del uso de esta tecnología se basa en la efectividad al analizar grandes volúmenes de datos.

En este sentido, el presente estudio ha sido desarrollado en una universidad española y tiene como objetivo estudiar el conocimiento y la preparación de los diferentes miembros de la comunidad universitaria (personal docente e investigador, personal de administración y servicios y estudiantes) para reconocer y, en consecuencia, reaccionar ante situaciones de ciberbullying. Para este fin, este trabajo se vale del uso de diferentes herramientas de inteligencia artificial generativa (GenAI), para analizar si pueden ser de ayuda para la detección de estos comportamientos. El objetivo se basa, por tanto, en evaluar la capacidad de detección de conductas relativas al ciberacoso tanto por parte de los miembros de la comunidad educativa, como por parte de diferentes herramientas de IA generativa establecer la efectividad de cada grupo en la identificación temprana de patrones de esta naturaleza.

1.2. La Inteligencia Artificial en la educación y el ciberacoso

La Inteligencia Artificial (IA) es una rama de la informática que se enfoca en el desarrollo de sistemas y máquinas capaces de realizar tareas que, de otra manera, requerirían de la inteligencia humana. En esencia, busca crear máquinas que puedan pensar, aprender y tomar decisiones de manera autónoma, imitando de alguna manera el comportamiento humano. Desde sus inicios, la IA ha abarcado una amplia gama de aplicaciones, desde el reconocimiento de voz y la visión por computadora hasta la predicción de patrones en grandes conjuntos de datos y la automatización de procesos complejos (Chaudhary et al., 2020). En particular, uno de los avances más significativos dentro del campo de la IA es la aparición de la IA generativa. Esta forma de IA se centra en la creación de modelos capaces de generar contenido nuevo, como imágenes, textos, música y vídeos, que en muchos casos se asemejan a los creados por humanos (Bandi, Adapa y Kuchi, 2023). En esencia, la IA generativa busca imitar la capacidad humana de crear y generar contenido original y significativo.

Los modelos de IA generativa funcionan mediante el uso de diferentes modelos de aprendizaje profundo (deep learning), como son las Redes Neuronales Generativas (GANs, por sus siglas en inglés) o las Redes Neuronales Recurrentes (RNNs). Estos modelos son entrenados con grandes cantidades de datos y son capaces de aprender patrones y características, que luego utilizan para generar nuevos datos que siguen esas mismas características aprendidas. En el contexto de la creación de contenido, la IA generativa ha tenido un impacto significativo en una variedad de campos, desde el arte y el diseño hasta la escritura y la música. De hecho, es la generación de texto la raíz de los modelos de IA generativa actuales y se considera que este es el área más avanzada (Barreto et al., 2023; Mohamadi et al., 2023). Sin duda, uno de los ejemplos más populares de la IA generativa generadora de texto son los Large Language Models (LLMs), modelos que pueden utilizarse para una enorme variedad de tareas para un amplio público no especializado (Hadi et al., 2023).

En el campo de la educación, los modelos de IA generativa tienen una amplia variedad de aplicaciones diferentes, y ya han sido considerados como una importante revolución en múltiples estudios diferentes (Lim et al., 2023; Malik et al., 2023; Sidiropoulos y Anagnostopoulos, 2024). Algunas de las ventajas del uso de estos modelos de lenguaje en el campo educativo son las tutorías personalizadas, el aprendizaje interactivo, la traducción automática de contenidos de texto, aplicaciones de audio y vídeo o herramientas para el aprendizaje adaptativo, que se ajusta al nivel actual de un alumno dentro de una determinada asignatura, para darle un aprendizaje ad hoc a sus necesidades. Por supuesto, la IA generativa no solo conlleva beneficios en la educación, sino también algunos problemas, como por ejemplo la falta de interacción humana, la aparición de sesgos o aspectos relacionados con la privacidad (Baidoo-Anu y Owusu Ansah, 2023). Como sucede con la mayoría de avances tecnológicos, la IA generativa también plantea numerosos desafíos éticos y sociales. Existen preocupaciones sobre la autenticidad y originalidad de las obras generadas por este tipo de IAs, así como sobre su potencial de uso malintencionado como, por ejemplo, la creación de noticias falsas o la generación de contenido acosador o amenazante (Baldassarre et al., 2023).

El papel que juega la IA y, particularmente, la IA generativa en relación al ciberacoso es doble y su rol está polarizado. Por un lado, tal y como señalan científicos sociales como Sameer Hinduja (Hinduja, 2023), la inteligencia artificial generativa posibilita la creación automática y la rápida difusión de mensajes, correos electrónicos, publicaciones o comentarios intimidantes o amenazantes en una diversidad de plataformas e interfaces. Sin embargo, por otro lado, puede ser utilizada como una herramienta efectiva para la propia detección del ciberacoso (Abarna et al., 2022), un hecho que se estudia pormenorizadamente en el presente trabajo. El hecho de que la IA permite analizar grandes cantidades de datos de manera rápida y precisa, permite abordar este problema de manera más eficiente que los métodos tradicionales de detección manual. La capacidad de procesar grandes volúmenes de datos es crucial para identificar patrones y tendencias en el comportamiento del ciberacoso. Adicionalmente, la IA puede aprender y adaptarse continuamente a medida que se le proporciona más información, pudiendo ajustarse a la aparición de nuevos patrones de ciberbullying con rapidez.

En concreto, las aplicaciones basadas en IA generativa, como por ejemplo chatbots como ChatGPT pueden desempeñar un papel importante como ayuda a los jóvenes que están sufriendo situaciones de acoso (Gabrielli et al., 2020; Koyuturk et al., 2023). Diferentes estudios demuestran que estos chatbots podrían convertirse en una herramienta para prevenir casos de acoso, a pesar de que todavía existen notables carencias en la detección de emociones, que su lenguaje no se adapta a la forma de hablar y escribir de los niños y que son demasiado predecibles (Lafrance St-Martin y Villeneuve, 2024). En el contexto de la detección del ciberacoso, la IA generativa en particular, puede utilizarse para crear modelos de lenguaje que identifiquen de forma automática el lenguaje ofensivo, amenazante o acosador. Un sistema de IA generativa puede entrenarse utilizando una gran cantidad de mensajes de ciberacoso conocidos, junto con ejemplos de comunicación on-line que no supongan situaciones de conflicto. A medida que el sistema procesa estos datos, aprende a distinguir entre las dos situaciones y puede identificar automáticamente mensajes que contengan señales de ciberacoso. Esto puede ser especialmente útil en plataformas de redes sociales y sistemas de mensajería, donde el ciberacoso puede ocurrir con frecuencia y a gran escala (Abarna et al., 2022; Schulenberg et al., 2023).

- En el presente trabajo, los autores estudian la capacidad de diferentes modelos de IA generativa como herramienta para la evaluación de mensajes que potencialmente puedan constituir situaciones de ciberbullying. Para este trabajo se han seleccionado cuatro de las aplicaciones más populares actualmente, como son ChatGPT-3.5, ChatGPT-4, Gemini y Llama2, los cuales se describen brevemente a continuación:
- ChatGPT-3.5: Es un modelo de lenguaje desarrollado por OpenAI, basado en la arquitectura GPT (Generative Pre-trained Transformer) que fue lanzado como una mejora del modelo GPT-3. En la actualidad es el modelo de acceso gratuito de la popular aplicación ChatGPT. Está disponible en la siguiente URL: https://chat.openai.com/.
- ChatGPT-4: También es un modelo de lenguaje desarrollado por OpenAI, en este caso en el modelo GPT-4. Su principal diferencia con respecto al GPT-3.5, además de ser una mejora en sí, es que puede trabajar con imágenes y otro tipo de contexto visual. Está disponible bajo suscripción de pago en la siguiente URL: https://chat.openai.com/.
- Gemini: Es un modelo de inteligencia artificial generativa desarrollado por Google, y busca ser la evolución de PaLM, el LLM en el que estaba basado Bard, el chatbot de IA generativa de Google. Está disponible en la siguiente URL: https://gemini.google.com/app.

LLaMa 2: Es un modelo de inteligencia artificial generativa de código abierto desarrollado por Meta para competir con OpenAI y Google en el sector de las IAs generativas. Actualmente no cuenta con una versión web chatbot oficial, pero está disponible en la siguiente URL: https://www.llama2.ai/. Estos son los modelos más populares de la actualidad y que están llamados a dominar el sector de los chatbots basados en IA generativa. Estos modelos han sido comparados en distintos estudios de la literatura por sus capacidades para desarrollar diferentes tareas, pero en este caso serán comparados por sus capacidades para la detección de mensajes que potencialmente puedan constituir situaciones de ciberacoso.

2. Protocolo experimental

2.1. Muestras y participantes

Este estudio se enmarca dentro de una tradición metodológica empírica y cuantitativa, seleccionada

específicamente por su capacidad para analizar patrones de percepción y detección de ciberacoso en redes sociales. Se han seleccionado y etiquetado 200 mensajes reales de redes sociodigitales (X, TikTok y YouTube) en función de su nivel de ciberacoso, utilizando una escala Likert de cinco niveles: no existe ciberacoso (1), ciberacoso leve (2), ciberacoso medio (3), ciberacoso alto (4) y ciberacoso muy alto (5). Esta metodología cuantitativa permite una comparación objetiva de las percepciones de diferentes actores de la comunidad educativa (estudiantes, personal docente e investigador, y personal administrativo) frente a un baseline establecido por un grupo de expertos. Asimismo, facilita la evaluación de la precisión de modelos de inteligencia artificial generativa (ChatGPT-3.5, ChatGPT-4, Gemini y LLaMa2) en la detección de patrones de ciberacoso. El enfoque cuantitativo es idóneo en este contexto porque permite obtener datos medibles y comparables sobre la capacidad de detección de cada grupo y de los modelos de IA, empleando métricas como el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Estas métricas ofrecen un marco de referencia para evaluar las diferencias en la identificación de ciberacoso.

- La recopilación de los mensajes utilizados fue llevada a cabo por los autores conjuntamente con un grupo de cinco expertos formado por profesionales de las siguientes áreas: criminología y derecho, psicología, sociología, pedagogía y educación. La selección de expertos en estas áreas garantiza un análisis integral del ciberacoso, abordando sus aspectos legales, psicológicos, sociales y educativos, y estableciendo un baseline de evaluación confiable. Estos campos proporcionan una perspectiva multidisciplinaria esencial para entender y clasificar el ciberacoso en toda su complejidad. A continuación, se describe brevemente el perfil de dichos expertos:
- Experta en criminología y derecho: doctora en derecho, magistrada y especialista en derecho penal y acoso.
- Experto en psicología: doctora en psicología y especialista en comportamiento criminal y en redes sociales.
- Experta en sociología: socióloga y especialista en comportamiento social.
- Experta en pedagogía: pedagoga, directora de la Unidad de Diversidad e Inclusión (UDI) de la universidad y especialista en inclusión y comportamientos de acoso dentro de la universidad.

  Experta en educación: doctora y especialista en educación.
- Una vez confeccionado el dataset con los 200 mensajes de los que iba a constar el estudio, cada uno de los expertos dió una puntuación del 1 (no existe ciberacoso) al 5 (ciberacoso muy alto) a cada uno de los mensajes, y el promedio de las respuestas de los cinco expertos, fue considerado como el baseline para los posteriores análisis del estudio. A continuación, se solicitó a los tres grupos de interés de una comunidad educativa universitaria que puntuásen los 200 mensajes en base a la misma escala que el grupo de expertos. Los tres grupos estudiados fueron:
- Grupo de alumnos: formado por 50 alumnos de distintos grados de las áreas de conocimiento de ciencias sociales, ciencias jurídicas, ciencias de la salud, ciencias de la educación y ciencia y tecnología.
- Grupo de personal docente e investigador (PDI): formado por 30 docentes de las mismas áreas de conocimiento.

Grupo de personal de administración y servicios (PAS): formado por 20 profesionales de las áreas de administración y servicios, entre los que se encuentran servicios de secretaría, atención al estudiante, coordinación académica, etc. La inclusión arbitraria de estudiantes, docentes y personal administrativo de la universidad permite capturar cómo cada grupo percibe y detecta el ciberacoso, según su experiencia y rol en el ámbito educativo, logrando así una comparación amplia y representativa de estas perspectivas en la comunidad universitaria. A cada uno de los miembros de los tres grupos anteriores (100 individuos en total), se le solicitó que ofrecieran una puntuación del 1 al 5 a cada uno de los mensajes, según la misma escala de Likert que los expertos. Tras esto, se calculó la respuesta promedio en cada uno de los tres grupos de interés, y se comparó con la respuesta del grupo de expertos.

Adicionalmente, y con el objetivo de establecer una comparación con los resultados obtenidos en los grupos anteriores, se pidió a las cuatro IAs generativas mencionadas en el apartado anterior que ofrecieran el mismo etiquetado del 1 al 5 a cada uno de los 200 mensajes. En los cuatro casos se ofreció a las inteligencias artificiales el mismo prompt, aunque cabe destacar que las salidas mostradas por las IAs no fueron las mismas: todas ellas daban el etiquetado solicitado para todos los mensajes, pero en algunos casos se incluían explicaciones adicionales a cada una de las etiquetas, como por ejemplo el tono detectado en el mensaje (ironía, confrontación, violencia, etc.). Esta información adicional se mostró con

los modelos de ChatGPT-3.5, ChatGPT-4 y Gemini. Además, Gemini generó recomendaciones para ampliar los conocimientos del usuario al respecto del ciberacoso, así como consejos a seguir en caso de haber detectado una de estas situaciones.

2.2. Instrumentos y procedimiento

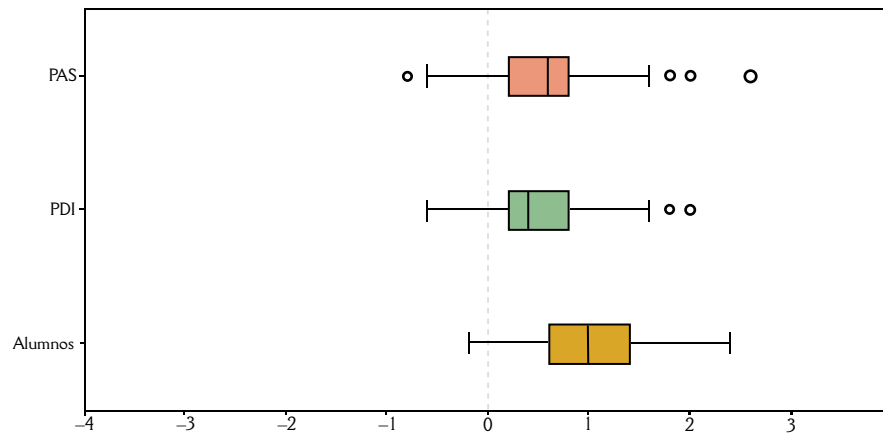
Las métricas de evaluación utilizadas para medir el error de clasificación de las IAs generativas y la comunidad educativa frente al baseline de los expertos son el RMSE (Root Mean Square Error – Error Cuadrático Medio) y el MAE (Mean Absolute Error – Error Absoluto Medio) (Chai y Draxler, 2014), que han sido ampliamente utilizadas en la literatura para este fin (Naser y Alavi, 2023). Estas dos métricas representan el promedio de las diferencias entre dos muestras de valores. Para la generación de los diagramas de tipo caja y bigotes que muestran las Figuras 1 y 2, se ha utilizado la diferencia entre la respuesta media de cada grupo de interés o IA y el grupo de expertos a cada una de las preguntas. Este tipo de gráfica permite una visualización clara de la dispersión de los datos de una muestra. Siguiendo el modelo de Tukey (Sidiropoulos y Anagnostopoulos, 2024), dentro de la caja sombreada en colores se encuentra el 50% de la distribución; el bigote izquierdo (respectivamente derecho) representa el 25% inferior (respectivamente superior) de la muestra. La línea vertical dentro de la caja representa la mediana de cada muestra, mientras que los círculos señalan los valores atípicos (outliers).

3. Resultados

3.1. La detección del ciberacoso dentro de una comunidad universitaria

Uno de los objetivos de este estudio es analizar las potenciales carencias y necesidades que los estudiantes y el personal de los centros educativos manifiesta ante situaciones de ciberacoso o conducta antisocial. En la Figura 1, se muestra la distribución de las diferencias obtenidas entre las respuestas de cada grupo de interés de la comunidad educativa respecto al baseline:

Figura 1: Distribución de las diferencias entre la comunidad educativa y los expertos.



Con carácter general, se detecta una diferencia positiva mayoritaria para los tres grupos estudiados. Esto indica que el etiquetado de los expertos es, en general, más alto que las valoraciones ofrecidas por el resto de grupos de la comunidad educativa universitaria. Los diagramas que representan el etiquetado ofrecido por el personal docente e investigador (PDI) y por el personal de administración y servicios (PAS) son notablemente similares, siendo la principal diferencia que la mediana de este último grupo ocupa una posición ligeramente mayor. Estas similitudes nos permiten analizar ambos diagramas de forma simultánea sin pérdida de rigurosidad. Como se ha mencionado previamente, el hecho de que los diagramas se sitúen notoriamente por encima del cero (representado por una línea vertical gris y discontínua), sugiere que el personal universitario (PDI y PAS), en comparación con los expertos, tiende a subestimar comentarios que potencialmente representan situaciones de abuso. En el caso de los alumnos, aproximadamente el 90% de las diferencias respecto a los

expertos toma valores positivos, situándose la distribución claramente desplazada hacia la derecha respecto a la del personal universitario. Los alumnos, por tanto, tienden a infraestimar notablemente aquellos comentarios que sí que representan ejemplos de acoso o actitudes perjudiciales y nocivas, difiriendo en algunas ocasiones en hasta más de dos puntos respecto a la valoración del grupo de expertos.

Tabla 1: Métricas de evaluación de las valoraciones de la comunidad educativa.

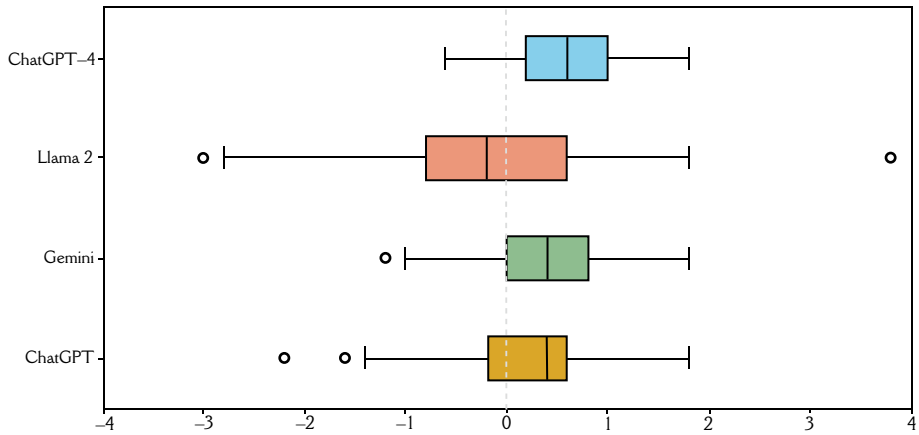
Grupo de interés	RMSE	MAE
Alumnos	1.18	1.03
Personal docente e investigador	0.69	0.55
Personal de administración y servicios	0.77	0.65
ChatGPT-3.5	0.70	0.57
Gemini	0.70	0.58
LLaMa2	1.10	0.85
ChatGPT-4	0.80	0.68

Adicionalmente, podemos observar en la Tabla 1 que las dos métricas de evaluación utilizadas van en la misma línea que la distribución observada en los diagramas de tipo caja y bigotes: el error promedio de los alumnos a la hora de etiquetar los mensajes es bastante mayor que en el caso de PDI y PAS, llegando los alumnos a obtener tasas de error de casi el doble que las obtenidas para el caso del personal docente.

3.2. Uso de IAs generativas como herramienta de ayuda en la detección del ciberacoso

En la actualidad, los diferentes modelos de IA generativa han demostrado ser herramientas eficaces a la hora de ayudar al ser humano en un amplio abanico de tareas. En este sentido, otro de los objetivos del presente trabajo es estudiar el potencial real de estos modelos a la hora de ayudar a los diferentes miembros de una comunidad educativa en la percepción, detección e incluso prevención, de situaciones de ciberacoso. De forma análoga al caso anterior, la Figura 2 representa la distribución de las diferencias obtenidas entre las respuestas ofrecidas por los cuatro modelos de IA generativa estudiados, con respecto al baseline:

Figura 2: Distribución de las diferencias entre las IAs generativas y los expertos.



En este caso, el diagrama de tipo caja y bigotes muestra una situación diferente. En primer lugar, si nos fijamos en el modelo LLaMa2, la distribución es uniforme en prácticamente todo el espectro, lo que ilustra que existen todo tipo de diferencias con respecto al grupo de expertos (es decir, se infravaloran situaciones de ciberacoso y se sobrevaloran situaciones en las que no existe ciberacoso). Adicionalmente, encontramos errores de hasta 3 unidades, con valores con un elevado grado de dispersión (alejados de la mediana). En segundo lugar, las distribuciones de los modelos Gemini y ChatGPT-4 son similares, aunque el primero de ellos está ligeramente desplazado hacia el centro del diagrama, indicando que, en general,

existen menores errores. Ambos dos modelos tienen una ligera tendencia a subestimar el ciberacoso, al estar la mayor parte de sus distribuciones sobre el eje positivo del diagrama.

En tercer lugar, los valores obtenidos por ChatGPT-3.5 están ligeramente más dispersos que en el caso de los modelos de Gemini y ChatGPT-4, pero están notablemente más centrados en el cero del diagrama, lo que nos indica que este modelo no tiene tanta tendencia a infravalorar las situaciones de ciberacoso como los tres modelos anteriores. También es importante destacar que ChatGPT-3.5 tiene una mayor dispersión de los datos (longitud de los bigotes del diagrama) que Gemini y ChatGPT-4. De forma análoga a lo sucedido con las métricas de evaluación de la Tabla 1 para los grupos de interés de la comunidad universitaria, en este caso podemos observar que los errores van en la misma línea de lo comentado: para el caso de LLaMa2, los valores de error son elevados, mientras que ChatGPT-3.5 y Gemini obtienen resultados muy similares a los del grupo del personal docente e investigador. ChatGPT-4 obtuvo errores ligeramente superiores, al igual que sucedía con el grupo del personal de administración y servicios.

Por último, llama especialmente la atención que el límite superior de los bigotes de los cuatro modelos coincide (en concreto, en el valor 1.8). Haciendo un análisis exhaustivo de la distribución de los errores, observamos que esto sólo sucede en los mensajes en los que los expertos dan una puntuación combinada de 2.8 (frente al 1 de los modelos de IA), pero no sucede en ningún caso en el que los expertos den una puntuación de 3.8 (y las IAs 2) o 4.8 (y las IAs 3), a excepción del dato atípico del modelo de LLaMa2. Este hecho es un indicador de que, a pesar de que los modelos de IA tienen errores, en las situaciones de ciberacoso claro (uso de palabras malsonantes, tono inconfundiblemente violento, etc.), las inteligencias artificiales tienen una mayor tasa de acierto que en situaciones algo más ambiguas (puntuaciones de 2 y 3 de los expertos). Estas situaciones generalmente coinciden con mensajes irónicos y uso de neolenguaje.

3. Limitaciones y recomendaciones

El presente trabajo ha buscado evaluar la capacidad de diferentes miembros de la comunidad educativa universitaria y diferentes inteligencias artificiales generativas a la hora de detectar y percibir comportamientos de acoso en redes sociales. Sin embargo, merece la pena mencionar ciertas limitaciones y consideraciones que deben ser consideradas al interpretar los resultados, así como para futuros trabajos. Por un lado, la muestra utilizada se compone únicamente de mensajes provenientes de tres redes sociales (X, TikTok y YouTube) y podría no reflejar todas las formas de ciberacoso que ocurren en otras plataformas o contextos. Además, aunque los mensajes fueron etiquetados por expertos en diferentes áreas, la percepción del ciberacoso puede variar según el contexto cultural o las experiencias individuales, lo que limita la generalización de los resultados a otros entornos.

Por otro lado, el uso de modelos de IA generativa de propósito general establece otra limitación relevante, ya que estos modelos no fueron específicamente entrenados para detectar ciberacoso. Esto puede afectar su precisión en comparación con modelos entrenados exclusivamente con datos de ciberacoso. Finalmente, el estudio se centra en una comunidad universitaria específica, lo que puede no ser representativo de otras comunidades educativas con características demográficas y culturales distintas. Con el fin de fortalecer la detección de ciberacoso y mejorar la aplicabilidad de los hallazgos, se recomienda que futuras investigaciones amplíen la variedad de plataformas y contextos de donde se recolectan los datos, para lograr capturar una mayor diversidad de casos y patrones de acoso. Además, sería valioso desarrollar y entrenar modelos de IA específicos para la detección de ciberacoso, lo cual podría mejorar la precisión y adaptabilidad de estas herramientas en entornos educativos. Asimismo, se sugiere profundizar en el análisis de otros estadísticos de dispersión y error, así como incluir estudios longitudinales para observar cómo las capacidades de detección del ciberacoso evolucionan con el tiempo en distintos grupos de interés. Por último, se recomienda integrar estas herramientas de IA en programas de formación y sensibilización para estudiantes, docentes y personal administrativo, con el fin de crear un entorno educativo más seguro y consciente, en el que todos los miembros de la comunidad puedan identificar y actuar frente al ciberacoso de manera efectiva.

4. Discusión y conclusiones

Con las consideraciones anteriormente descritas, el estudio nos permite inferir varias conclusiones, que se detallan a continuación. En primer lugar, con un MAE inferior al resto de grupos, ha quedado demostrado que los docentes son el grupo de la comunidad educativa estudiada con mejor preparación para detectar

situaciones de ciberacoso, aunque su capacidad tiene todavía un notable potencial de mejora. El grupo de personal de administración y servicios obtuvo ligeramente peores resultados que el personal docente. En general, esto indica que sería positivo para la comunidad educativa que estos dos grupos mejorasen en sus capacidades de detección de situaciones de ciberacoso, para así poder dar un mejor servicio y protección a aquellos individuos que son víctimas de este tipo de situaciones. Es importante destacar que el tiempo de detección de una situación de acoso es vital a la hora de evitar más sufrimiento para la víctima y/o potenciales situaciones futuras más graves.

En segundo lugar, el grupo de alumnos es el grupo de la comunidad educativa que más infravalora las situaciones de ciberacoso (valorando en bastantes casos con puntuaciones de 2 unidades inferiores a las valoraciones ofrecidas por el grupo de expertos). El elevado RMSE de este grupo sugiere una alta variabilidad en la precisión al clasificar el ciberacoso, lo cual puede llevar a subestimaciones significativas de casos reales de acoso. Este resultado implica que los estudiantes podrían no identificar adecuadamente el ciberacoso en su entorno, lo que podría reducir su disposición a intervenir o informar sobre tales situaciones. Se pone de manifiesto que es necesario mejorar la formación y concienciación del alumnado ante estas situaciones de ciberacoso, ya que, en muchos casos, son los propios alumnos los que tienen información más cercana sobre situaciones de ciberacoso que se están produciendo que el propio personal universitario (PDI y PAS). En tercer lugar, este estudio ha puesto en evidencia que algunos de los modelos de inteligencia artificial generativa populares hoy en día pueden ser utilizados como una herramienta de ayuda a la hora de concienciar, ayudar y dar soporte a los miembros de la comunidad educativa ante la aparición de situaciones de ciberacoso y a la prevención de las mismas.

La variabilidad observada a través de estadísticos como la mediana y los percentiles en los diagramas de caja y bigotes ayuda a identificar hasta qué punto los diferentes grupos, y las IAs, se apartan de la evaluación experta. Los estudiantes presentan la mayor dispersión, lo que sugiere una comprensión inconsistente del ciberacoso, mientras que el personal docente y las IAs muestran menores niveles de dispersión, indicando mayor consistencia. Esta dispersión tiene implicaciones en la práctica educativa: una baja variabilidad en la capacidad de detectar ciberacoso permitiría una respuesta uniforme y rápida en casos de acoso, reduciendo la probabilidad de que estos casos sean pasados por alto.

Sin embargo, también es importante comentar que, de forma general, estos modelos de IA cometen errores (en el mejor de los casos, ofrecen errores muy similares al del mejor grupo de humanos estudiado, el personal docente e investigador), lo que destaca que estos modelos todavía tienen un importante potencial de mejora. Así mismo, es importante comentar que estos modelos estudiados son modelos de propósito general, es decir, no han sido diseñados para desarrollar la tarea de detección de situaciones de ciberacoso de forma específica, por lo que un modelo de inteligencia artificial entrenado con una gran cantidad de datos de ciberacoso etiquetados adecuadamente, ofrecería mejores resultados, tal y como prueban algunos ejemplos recientes que se pueden encontrar en la literatura (Iwendi et al., 2023; Kargutkar y Chitre, 2020). Cabe destacar que la ventaja que ofrecen las IAs generativas como ChatGPT es que son muy populares, ya han sido ampliamente aceptadas por la sociedad y están siendo utilizadas para un gran abanico de tareas, por lo que son mucho más accesibles para los miembros de la comunidad educativa que otro tipo de tecnologías basadas en inteligencia artificial.

En cuarto lugar, es importante destacar que los modelos de IA generativa no solo ayudan a la detección del ciberacoso, sino que también pueden ser de gran ayuda en cuanto a la prevención y concienciación (Aggarwal y Gaur, 2024; Tesfagergish y Damaševičius, 2024), tanto del alumnado como del personal educativo (PDI y PAS). Distintos estudios han demostrado este hecho, pero también puede utilizarse para lo contrario: los modelos de IA generativa pueden cometer errores que faciliten la distribución y viralización de información falsa, el acoso masivo en redes sociales mediante bots, etc. (Sidiropoulos y Anagnostopoulos, 2024). Se destaca también que los errores de clasificación de las IAs generativas pueden provocar pérdida de confianza en el sistema: los errores positivos (infravaloración de situaciones frente a los expertos), pueden facilitar que aparezcan situaciones de ciberacoso, mientras que los errores negativos (sobrevaloración de situaciones frente a los expertos), pueden provocar censura o situaciones de injusticia.

El estudio realizado resalta la importancia de las herramientas de GenAI como método para detectar patrones de ciberacoso, lo que puede ser útil para reducir el acoso en redes sociales, tanto internas como externas al centro educativo, dentro de la comunidad docente. Aunque es necesario continuar mejorando

el entrenamiento de estas herramientas para lograr una detección más precisa de estos comportamientos, pueden actuar como una alerta inicial para responder rápidamente en situaciones donde el tiempo de reacción es crucial para evitar daños a los estudiantes. Podemos concluir que las IA generativas pueden facilitar la detección temprana del ciberacoso, permitiendo una intervención rápida por parte de las unidades responsables dentro del centro educativo.

Además, el uso de estas herramientas puede contribuir a la formación y capacitación del personal educativo para detectar estos casos de manera clara y eficiente. Por último, los estudiantes también pueden recibir formación en ciberacoso basada en los resultados obtenidos mediante el análisis de las IA generativas, así como en el uso de herramientas de alerta generadas por estas aplicaciones. Los resultados de dispersión y error sugieren que una IA generativa bien entrenada podría ser una herramienta efectiva para complementar la detección de ciberacoso. Sin embargo, el estudio también muestra limitaciones en la IA, especialmente en casos ambiguos. Esto implica que, aunque la IA puede servir de soporte, es fundamental que el personal educativo mantenga una capacitación continua para interpretar y verificar sus resultados. Con sus limitaciones, este estudio aporta conocimiento relevante a la escasa literatura existente sobre el rol protector que pueden desempeñar el entorno educativo en la moderación del impacto del acoso y ciberacoso entre jóvenes. Integrar tanto la IA como estrategias de sensibilización y capacitación en la comunidad educativa podría mejorar la detección de ciberacoso, proporcionando una protección más robusta y una mejor convivencia en entornos escolares.

Apoyos

No aplica.

Referencias

Abarna, S., Sheeba, J. I., Jayasrilakshmi, S. y Devaneyan, S. P. (2022). Identification of cyber harassment and intention of target users on social media platforms. *Engineering Applications of Artificial Intelligence*, *115*, 105283. https://doi.org/10.1016/j.engappai.2022.105283

Aggarwal, A. y Gaur, S. (2024). Applying Artificial Intelligence to Explore Online Harassment and Cyberbullying Prevention. En S. Ponnusamy, V. Bora, P. M. Daigavane, y S. S. Wazalwar (Eds.), *Impact of AI on Advancing Women's Safety* (pp. 104-120). IGI Global. https://doi.org/10.4018/979-8-3693-2679-4.ch007

Azeez, N. A., Idiakose, S. O., Onyema, C. J. y Van Der Vyver, C. (2021). Cyberbullying Detection in Social Networks: Artificial Intelligence Approach. *Journal of Cyber Security and Mobility*, *10*(4), 745-774. https://doi.org/10.13052/jcsm2245-1439.1046

Baidoo-Anu, D. y Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, *7*(1), 52-62. https://doi.org/10.61969/jai.1337500

Baldassarre, M. T., Caivano, D., Fernandez Nieto, B., Gigante, D. y Ragone, A. (2023). The Social Impact of Generative AI: An Analysis on ChatGPT. En *Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (pp. 363-373). Association for Computing Machinery. https://doi.org/10.1145/3582515.3609555

Bandi, A., Adapa, P. V. S. R. y Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. *Future Internet*, *15*(8), 260. https://doi.org/10.3390/fi15080260

Barlett, C. P. (2015). Anonymously Hurting Others Online: The Effect of Anonymity on Cyberbullying Frequency. *Psychology of Popular Media Culture*, *4*(2), 70-79. https://doi.org/10.1037/a0034335

Barreto, F., Moharkar, L., Shirodkar, M., Sarode, V., Gonsalves, S. y Johns, A. (2023). Generative Artificial Intelligence: Opportunities and Challenges of Large Language Models. En V. E. Balas, V. B. Semwal, y A. Khandare (Eds.), *Intelligent Computing and Networking* (pp. 545-553). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3177-4_41

Burger, C., Strohmeier, D. y Kollerová, L. (2022). Teachers Can Make a Difference in Bullying: Effects of Teacher Interventions on Students' Adoption of Bully, Victim, Bully-Victim or Defender Roles across Time. *Journal of Youth and Adolescence*, *51*(12), 2312-2327. https://doi.org/10.1007/s10964-022-01674-6

Campbell, M., Whiteford, C. y Hooijer, J. (2019). Teachers' and parents' understanding of traditional and cyberbullying. *Journal of School Violence*, *18*(3), 388-402. https://doi.org/10.1080/15388220.2018.1507826

Cerezo, F. (2009). Bullying: análisis de la situación en las aulas españolas. *International Journal of Psychology and Psychological Therapy*, *9*(3), 383-394. https://bit.ly/42UCz0c

Chai, T. y Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chaudhary, S., Yadav, S., Kushwaha, S. y Shahi, S. R. P. (2020). A Brief Review of Machine Learning and its Applications. *Samriddhi: A Journal of Physical Sciences, Engineering and Technology*, *12*(SUP 1), 218-223. https://www.smsjournals.com/index.php/SAMRIDDHI/article/view/1941

Cortés, A. F. M., De los Río, O. L. H. y Pérez, A. S. (2019). Factores de riesgo y factores protectores relacionados con el ciberbullying entre adolescentes: una revisión sistemática. *Papeles del psicólogo*, *40*(2), 109-124. https://doi.org/10.23923/pap.psicol2019.2899

Crothers, L. M. y Levinson, E. M. (2004). Assessment of Bullying: A Review of Methods and Instruments. *Journal of Counseling & Development*, *82*(4), 496-503. https://doi.org/10.1002/j.1556-6678.2004.tb00338.x

Egeberg, G., Thorvaldsen, S., Rønning, J. A. y Elstad, E. (2016). Digital Expectations and Experiences in Education. En *The Impact of Cyberbullying and Cyber Harassment on Academic Achievement* (pp. 183-204). Brill. https://doi.org/10.1007/9789463006484_012

Gabrielli, S., Rizzi, S., Carbone, S. y Donisi, V. (2020). A Chatbot-Based Coaching Intervention for Adolescents to Promote Life Skills: Pilot Study. *JMIR Human Factors*, *7*(1), e16762. https://doi.org/10.2196/16762

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., et al. (2023). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *Authorea Preprints*. https://doi.org/10.36227/techrxiv.23589741.v1

Hinduja, S. (2023, Mayo 10). *Generative AI as a Vector for Harassment and Harm*. Cyberbullying Research Center. https://bit.ly/42S9Mtf

Hinduja, S. y Patchin, J. W. (2008). Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behavior*, *29*(2), 129-156. https://doi.org/10.1080/01639620701457816

Houbre, B., Tarquinio, C., Thuillier, I. y Hergott, E. (2006). Bullying among students and its consequences on health. *European Journal of Psychology of Education*, *21*(2), 183-208. https://doi.org/10.1007/BF03173576

Iwendi, C., Srivastava, G., Khan, S. y Maddikunta, P. K. R. (2023). Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*, *29*(3), 1839-1852. https://doi.org/10.1007/s00530-020-00701-5

Kargutkar, S. M. y Chitre, V. (2020). A Study of Cyberbullying Detection Using Machine Learning Techniques. En *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 734-739). IEEE. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000137

Kowalski, R. M. y Limber, S. P. (2007). Electronic Bullying Among Middle School Students. *Journal of Adolescent Health*, *41*(6, Supplement), S22-S30. https://doi.org/10.1016/j.jadohealth.2007.08.017

Koyuturk, C., Yavari, M., Theophilou, E., Bursic, S., Donabauer, G., Telari, A., et al. (2023). Developing effective educational Chatbots with ChatGPT prompts: insights from preliminary tests in a case study on social media literacy (with appendix). *arXiv preprint arXiv:2306.10645*. https://doi.org/10.48550/arXiv.2306.10645

Lafrance St-Martin, L. I. y Villeneuve, S. (2024). The uses of chatbots in the context of children and teenagers bullying: a systematic literature review. *Cogent Education*, *11*(1), 2312032. https://doi.org/10.1080/2331186X.2024.2312032

Landstedt, E. y Persson, S. (2014). Bullying, cyberbullying, and mental health in young people. *Scandinavian Journal of Public Health*, *42*(4), 393-399. https://doi.org/10.1177/1403494814525004

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I. y Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, *21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lucas-Molina, B., Pérez-Albéniz, A., Solbes-Canales, I., Ortuño-Sierra, J. y Fonseca-Pedrero, E. (2022). Bullying, Cyberbullying and Mental Health: The Role of Student Connectedness as a School Protective Factor. *Psychosocial Intervention*, *31*(1), 33-41. https://doi.org/10.5093/PI2022A1

Malik, T., Hughes, L., Dwivedi, Y. K. y Dettmer, S. (2023). Exploring the Transformative Impact of Generative AI on Higher Education. En M. Janssen, L. Pinheiro, R. Matheus, F. Frankenberger, Y. K. Dwivedi, I. O. Pappas, y M. Mäntymäki (Eds.), *New Sustainable Horizons in Artificial Intelligence and Digital Solutions* (pp. 69-77). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-50040-4_6

Mohamadi, S., Mujtaba, G., Le, N., Doretto, G. y Adjeroh, D. A. (2023). ChatGPT in the Age of Generative AI and Large Language Models: A Concise Survey. *arXiv preprint arXiv:2307.04251*. https://doi.org/10.48550/arXiv.2307.04251

Muzamil, M. y Shah, G. (2016). Cyberbullying and Self-Perceptions of Students Associated With Their Academic Performance. *International Journal of Education and Development Using ICT*, *12*(3), 79-92. https://bit.ly/3TfWvY3

Naser, M. Z. y Alavi, A. H. (2023). Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Architecture, Structures and Construction*, *3*(4), 499-517. https://doi.org/10.1007/s44150-021-00015-8

Noblia, V., Renato, A. C. y Gershanik, T. (2022). Anonimato, pseudonimia y delitos en las redes sociales: una propuesta multidimensional de la lingüística forense para la identificación de autoría. *Revista Latinoamericana de Estudios Del Discurso*, *22*(1), 122-142. https://doi.org/10.35956/v.22.n1.2022.p.122-142

Pichel, R., Feijóo, S., Isorna, M., Varela, J. y Rial, A. (2022). Analysis of the relationship between school bullying, cyberbullying, and substance use. *Children and Youth Services Review*, *134*, 106369. https://doi.org/10.1016/j.childyouth.2022.106369

Reynoso, T. M., González, B. M. y López, A. S. (2021). Ciberbullying, brecha digital y habilidades digitales para ciberconvivencia: descripción en estudiantes de bachillerato. *Voces de la educación*, *6*(12), 22-44. https://www.revista.vocesdelaeducacion.com.mx/index.php/voces/article/view/386

Schneider, S. K., O'donnell, L., Stueve, A. y Coulter, R. W. S. (2012). Cyberbullying, School Bullying, and Psychological Distress: A Regional Census of High School Students. *American Journal of Public Health*, *102*(1), 171-177. https://doi.org/10.2105/AJPH.2011.300308

Schulenberg, K., Li, L., Freeman, G., Zamanifard, S. y McNeese, N. J. (2023). Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. En A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, y M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581090

Sidiropoulos, D. y Anagnostopoulos, C.-N. (2024). Applications, challenges and ethical issues of AI and ChatGPT in education. *arXiv preprint arXiv:2402.07907*. https://doi.org/10.48550/arXiv.2402.07907

Tesfagergish, S. G. y Damaševičius, R. (2024). Explainable Artificial Intelligence for Combating Cyberbullying. En K. K. Patel, K. C. Santosh, A. Patel, y A. Ghosh (Eds.), *Soft Computing and Its Engineering Applications* (pp. 54-67). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53731-8_5

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277-287. https://doi.org/10.1016/j.chb.2009.11.014

Tukey, J. W. (1993). *Exploratory Data Analysis: Past*, *Present and Future*. Defense Technical Information Center. https://bit.ly/3SZ1nj6

Ventura, P. D. B., Rodríguez-García, A. M. y Reche, J. M. S. (2018). Incidencia del ciberbullying en adolescentes de 11 a 17 años en Portugal. *Edutec. Revista Electrónica de Tecnología Educativa*, (64), 82-98. https://doi.org/10.21556/edutec.2018.64.1029

Wang, J., Iannotti, R. J. y Nansel, T. R. (2009). School Bullying Among Adolescents in the United States: Physical, Verbal, Relational, and Cyber. *Journal of Adolescent Health*, *45*(4), 368-375. https://doi.org/10.1016/j.jadohealth.2009.03.021

Ybarra, M. L. y Mitchell, K. J. (2004). Youth engaging in online harassment: associations with caregiver–child relationships, Internet use, and personal characteristics. *Journal of Adolescence*, *27*(3), 319-336. https://doi.org/10.1016/j.adolescence.2004.03.007

Yousef, W. S. M. y Bellamy, A. (2015). The Impact of Cyberbullying on the Self-Esteem and Academic Functioning of Arab American Middle and High School Students. *Electronic Journal of Research in Educational Psychology*, *13*(3), 463-482. https://doi.org/10.14204/ejrep.37.15011