Artículos

Extracción de Palabras Clave de Ciberacoso de Textos Breves: un Enfoque de Aprendizaje Automático

Extracting Cyberbullying Keywords from Short Texts: A Machine Learning Approach

William Hermel Astudillo Quituisaca
Laboratorio de Investigación y Desarrollo en Informática (LIDI), Ecuador
wastudillomsn@es.uazuay.edu.ec

Priscila Cedillo
Universidad de Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Marcos Orellana
Laboratorio de Investigación y Desarrollo en Informática (LIDI), Ecuador
marore@uazuay.edu.ec

Revista Tecnológica ESPOL - RTE
vol. 36, núm. 1, Esp. p. 25 - 38, 2024
Escuela Superior Politécnica del Litoral, Ecuador
ISSN: 0257-1749
ISSN-E: 1390-3659
Periodicidad: Semestral
rte@espol.edu.ec

Recepción: 12 Julio 2024
Aprobación: 02 Octubre 2024

DOI: https://doi.org/10.37815/rte.v36nE1.1207

URL: https://portal.amelica.org/ameli/journal/844/8445128002/

Resumen: El ciberacoso impacta negativamente a la sociedad debido a las consecuencias que sufren las víctimas, acosadores y espectadores. El acceso generalizado a Internet y redes sociales, especialmente entre jóvenes sin herramientas para enfrentar estas situaciones, hace necesaria una formación social que mitigue los efectos del ciberacoso. Este estudio busca contribuir a esa formación mediante la creación de guiones para cápsulas educativas. Para ello, se desarrolló un modelo que automatiza la búsqueda y extracción de datos de la red social X utilizando Python y Selenium Web Driver. Tras un proceso de pre-procesamiento de textos utilizando técnicas de Procesamiento de Lenguaje Natural, se aplicó el modelo de Asignación Latente de Dirichlet (LDA) para identificar las palabras clave. Finalmente, se utilizó el modelo pre-entrenado "text-davinci-003" a través de la API de la empresa OpenAI para generar el contenido de las cápsulas educativas, asignando un contexto y utilizando las palabras clave identificadas. El resultado de esta investigación propuesta es la generación de un guion que contiene temas de educación y prevención del acoso y ciberacoso. Para garantizar una confiabilidad del texto generado por el modelo pre-entrenado generativo, se evaluó con un experto en la materia mediante el enfoque de Meta-Pregunta-Respuesta (GQM), lo que valida su potencial en la generación de contenido educativo en la lucha contra el ciberacoso.

Palabras clave: Acoso, Asignación Latente de Dirichlet, Ciberacoso, Inteligencia artificial, Modelos de lenguaje.

Abstract: Cyberbullying has a negative impact on society due to the consequences suffered by victims, bullies, and bystanders. Widespread access to the internet and social networks, especially among young people without the tools to

deal with these situations, makes social education necessary to mitigate the effects of cyberbullying. This study seeks to contribute to this training through the creation of scripts for educational capsules. To this end, a model was developed that automates the search and extraction of data from the social network X using Python and Selenium Web Driver. After a text preprocessing process using Natural Language Processing techniques, the Latent Dirichlet Assignment (LDA) model was applied to identify keywords. Finally, the pre-trained model "text-davinci-003" was used through the OpenAI API to generate the content of the educational capsules, assigning a context and using the identified keywords. The outcome of this proposed research is the generation of a script that includes topics on education and the prevention of bullying and cyberbullying. To ensure the reliability of the text generated by the pre-trained generative model, it was evaluated by an expert in the field using the Goal-Question-Metric (GQM) approach, which validates its potential in generating educational content in the fight against cyberbullying.

Keywords: Artificial Intelligence, Bullying, Cyberbullying, Latent Dirichlet Assignment, Language models.

0

El uso de las Tecnologías de la Información y Comunicación (TIC) se ha integrado profundamente en la vida diaria, especialmente entre los jóvenes. En Ecuador, el 92% de la población mayor a cinco años utiliza redes sociales (Peña y Herrera, 2021). Si bien las TIC ofrecen múltiples beneficios, también facilitan comportamientos negativos como el ciberacoso. Garaigordobil (2014) define el ciberacoso como una forma de acoso que utiliza las TIC, y el acoso escolar tradicional como una violencia reiterada entre iguales, donde uno o más agresores ejercen poder sobre una víctima para causarle daño. En estas dinámicas se identifican tres roles: i) el agresor, la o las personas que practican el acoso; ii) el espectador, la o las personas que presencian situaciones de acoso; y iii) la víctima, la o las personas acosadas. Este incremento en el uso de las TIC y esta forma de acoso resaltan la importancia de promover una educación apropiada en el uso de la tecnología, especialmente entre los niños y jóvenes, quienes son los más expuestos a estos riesgos.

El acoso escolar no solo afecta la salud mental, física y psicológica de las víctimas, sino que también aumenta el riesgo de deserción escolar. Las víctimas experimentan un incremento en la ansiedad y la depresión, mientras que los espectadores pueden desarrollar miedo, sumisión, desensibilización e incluso interiorizar conductas antisociales y delictivas como medio para obtener lo que desean, además de experimentar sentimientos de culpa (Lugones Botell y Ramírez Bermúdez, 2017). Bajo este contexto, el ciberacoso se ha convertido en un problema global con graves consecuencias sociales, por lo que su mitigación se ha convertido en una necesidad urgente. Herramientas como los programas y campañas de prevención han contribuido en la reducción de estudiantes víctimas de acoso escolar y ciberacoso (Salmivalli et al., 2021). Sin embargo, una educación adecuada tanto de estudiantes como de padres en temas relacionados con el acoso y el fortalecimiento de la autoestima es importante para que modelos de prevención de acoso y ciberacoso sean efectivos.

Una forma de promover el aprendizaje, especialmente en niños y jóvenes, es a través de nuevas metodologías como el micro aprendizaje, donde la atención de la audiencia es captada mediante contenidos breves y fáciles de consumir en periodos cortos. Por lo tanto, una herramienta popular en este enfoque son las cápsulas educativas, que combinan las TIC con la generación de contenidos digitales educativos (Vidal Ledo et al., 2019). Dado que los contenidos son concisos, es fundamental que sean relevantes, precisos y presenten la información de manera clara y comprensible (Kamilali y Sofianopoulou, 2015).

Bajo este contexto, X (anteriormente Twitter) se ha consolidado como un espacio ideal para la expresión de opiniones, comentarios y la difusión de contenido digital, gracias a su característica de "hilos" que permite interconectar publicaciones en base a temas o tópicos similares (Guallar y Traver, 2020). Además, Arazzi et al. (2023) señalan que el lenguaje en redes sociales está en constante evolución y que existe una relación entre la forma en que las personas se expresan en estas plataformas y el impacto que generan en las comunidades digitales. Siendo así, esta dinámica ha convertido a la plataforma en un valioso recurso para la investigación en diversos temas sociales, facilitando el filtrado y difusión de contenidos. El formato predominante de datos en X, basado en texto plano, simplifica el análisis de contenido. Sin embargo, la plataforma ofrecía una interfaz de programación de aplicaciones (API, por sus siglas en inglés) gratuita para la búsqueda y extracción de contenido; ahora este acceso es de pago, lo que presenta limitaciones a las investigaciones que solían utilizar datos de X a gran escala.

Por otra parte, la ciencia de datos busca generar conocimiento a través de la extracción y análisis de información, utilizando técnicas estadísticas y matemáticas que facilitan la interpretación precisa de los datos y resultados (Orellana et al., 2023). Esto la convierte en una herramienta fundamental para la toma de decisiones informadas (Azuela y Ayala, 2019).

El Procesamiento del Lenguaje Natural (PLN) se enfoca en proporcionar métodos para analizar, modelar y comprender el lenguaje humano. Esta disciplina nos permite aprovechar la vasta cantidad de textos disponibles en el entorno digital para extraer información valiosa y enriquecer nuestra comprensión de diversos temas. Dentro del campo del PLN, existen técnicas como la Asignación Latente de Dirichlet (LDA, por sus siglas en inglés), una técnica destacada para el análisis de grandes volúmenes de texto. Este modelo probabilístico transforma un corpus de texto en un conjunto reducido de temas o tópicos, cada uno representado por una distribución de palabras. Esta técnica asume que los documentos son una mezcla de tópicos que generan palabras según su distribución de probabilidad. Esta representación procesa eficientemente grandes cantidades de datos, preservando las relaciones estadísticas clave. LDA ha sido utilizada en varias tareas, como clasificación de textos, modelado de tópicos, detección de anomalías textuales, análisis de similitud, extracción de palabras clave, entre otras (Blei et al., 2003).

En este trabajo, se propone utilizar textos de la red social X relacionados con el acoso escolar y el ciberacoso para analizar las palabras y frases más relevantes. A partir de este análisis, se busca contribuir a la generación de guiones para combatir estas problemáticas empleando procesos y herramientas que faciliten la

búsqueda y generación de información. Para ello, se utiliza como base la metodología definida en el proceso transversal estándar para la minería de datos (CRISP-DM, por sus siglas en inglés), considerada un estándar en proyectos de minería de datos. Esta metodología destaca por su simplicidad, estructura, confiabilidad y amplia aplicación en diversos modelos de procesos en varias áreas de conocimiento (Mancilla-Vela et al., 2020).

Adicionalmente, el presente documento se estructura de la siguiente manera: la Sección 2 presenta los trabajos relacionados, la Sección 3 detalla los materiales y métodos utilizados, la Sección 4 expone los resultados del estudio y la Sección 5 plantea las conclusiones y trabajos futuros.

Trabajos relacionados

La red social Twitter o ahora llamada X es una plataforma de gran importancia para la investigación y el abordaje de diversas problemáticas, incluyendo el acoso escolar y el ciberacoso. Numerosos estudios se han centrado en examinar la información de X relacionada con estas temáticas, ya sea obtenida a través de su API, o de conjuntos de datos disponibles en repositorios digitales. Estos análisis brindan una mejor comprensión de las dinámicas del acoso escolar y el ciberacoso, además de identificar patrones de comportamiento y desarrollar estrategias de prevención e intervención efectivas.

Es así como Chen et al. (2022) evaluaron el potencial de X como fuente de datos en la investigación, analizando aspectos como el acceso, costo, habilidades requeridas y calidad de los datos. Concluyeron que, si bien los datos de X son relevantes para diversas áreas de estudio, su obtención puede requerir conocimientos de programación que no todos los investigadores poseen, especialmente con la transición hacia una API de pago.

Alim (2015) investigó tweets relacionados con el ciberacoso, analizando características como el uso de hashtags, URLs, antigüedad de perfiles y número de seguidores. El estudio reveló que algunos tweets ofrecían consejos sobre cómo enfrentar el ciberacoso, lo cual podría ser valioso para prevenir y combatir este problema. De igual manera, Guallar y Traver (2020) investigaron la curación de contenidos digitales en X, analizando cómo los usuarios buscan, organizan y filtran información en la plataforma. Su estudio resaltó la utilidad de X, particularmente a través de hilos, para acceder y enriquecer información de diversas temáticas, agregando valor al contenido original.

Por otra parte, Sanchez y Kumar (2011) realizaron un análisis de sentimientos en tweets relacionados con el acoso escolar, utilizaron la API de X para acceder a los datos y crear un conjunto de

entrenamiento para un clasificador Naïve Bayes. No obstante, encontraron dificultades debido a las medidas de protección de X, que eliminaba tweets antes de que pudieran ser analizados, lo que limitó el alcance de su estudio.

Bayari y Bensefia (2021) llevaron a cabo un estudio sobre la detección automática de ciberacoso basado en el análisis de contenido textual. Para ello, exploraron diversas técnicas de minería de textos, como la Bolsa de Palabras (BoW, por sus siglas en inglés) y las Características Léxicas y Sintácticas (LSF, por sus siglas en inglés). Utilizaron la red social X como fuente principal de datos y resaltaron la importancia de considerar el idioma en las reglas y estructuras de análisis, ya que este influye significativamente en la forma en que se manifiesta el ciberacoso.

Un enfoque distinto es presentado por Vázquez et al. (2017), donde realizaron una revisión exhaustiva de modelos empleados en la generación automática de diálogos, centrándose en la generación de lenguaje natural. Examinaron modelos bayesianos, probabilísticos, estocásticos y de redes neuronales, y concluyeron que los modelos basados en redes neuronales generan respuestas más adecuadas y coherentes en el contexto de un diálogo. Bajo este contexto, Fatima et al. (2022) llevaron a cabo una revisión exhaustiva sobre la generación de texto utilizando modelos de redes neuronales profundas. En su estudio, resaltaron la relevancia de GPT-3, un modelo pre-entrenado con una vasta cantidad de textos y parámetros, lo que le permite generar textos de alta calidad en diversas tareas.

De esta manera, el presente estudio propone un novedoso enfoque para la prevención del acoso escolar y el ciberacoso, basado en el análisis de grandes volúmenes de datos textuales provenientes de la red social X. A través de técnicas de minería de texto, se busca identificar patrones lingüísticos y semánticos recurrentes en las conversaciones relacionadas con estas problemáticas, con la finalidad de generar guiones para cápsulas educativas en temas de educación y prevención del ciberacoso a través de modelos pre-entrenados generativos.

Materiales y Métodos

Para el desarrollo de las cápsulas de aprendizaje para el combate y la prevención del ciberacoso se ha utilizado como guía la metodología de la empresa IBM definida como el proceso transversal estándar para la minería de datos (CRISP-DM). Esta metodología consta de seis fases: i) comprensión del negocio, ii) comprensión de los datos, iii) preparación de los datos, iv) modelado, v) evaluación, y vi) despliegue. El ciclo de vida tradicional de un proyecto con esta metodología se describe en la Figura 1.

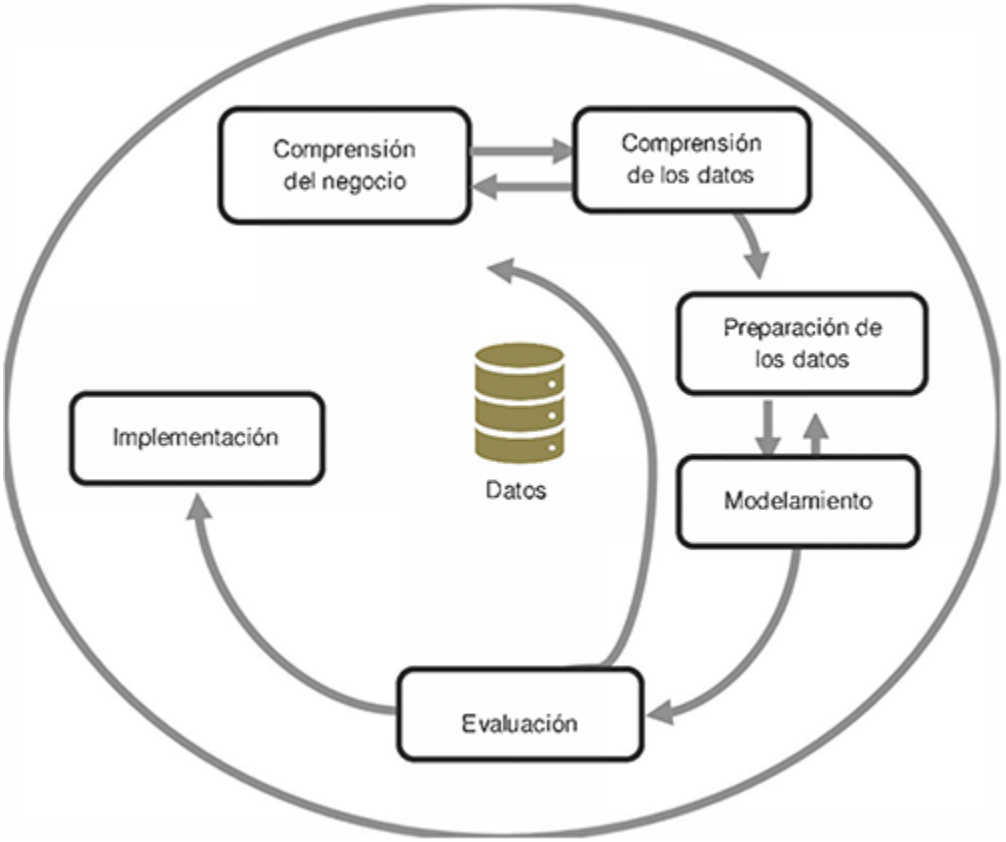

Figura 1

Ciclo de vida de proyecto con CRISP-DM

Nota. Adaptado del proceso de la Metodología CRISP-DM por Mancilla-Vela et al (2020).

Adicionalmente, se adoptó la especificación del Metamodelo de Ingeniería de Procesos de Sistemas 2.0 (SPEM, por sus siglas en inglés) para representar las fases de la metodología basada en el ciclo de vida de CRISP-DM. Por lo tanto, la metodología fue representada en seis fases: i) definición de criterios de búsqueda, ii) extracción de datos, iii) pre-procesamiento de datos, iv) extracción de palabras clave, v) generación del guion y vi) validación. La Figura 2 ilustra la representación de la metodología.

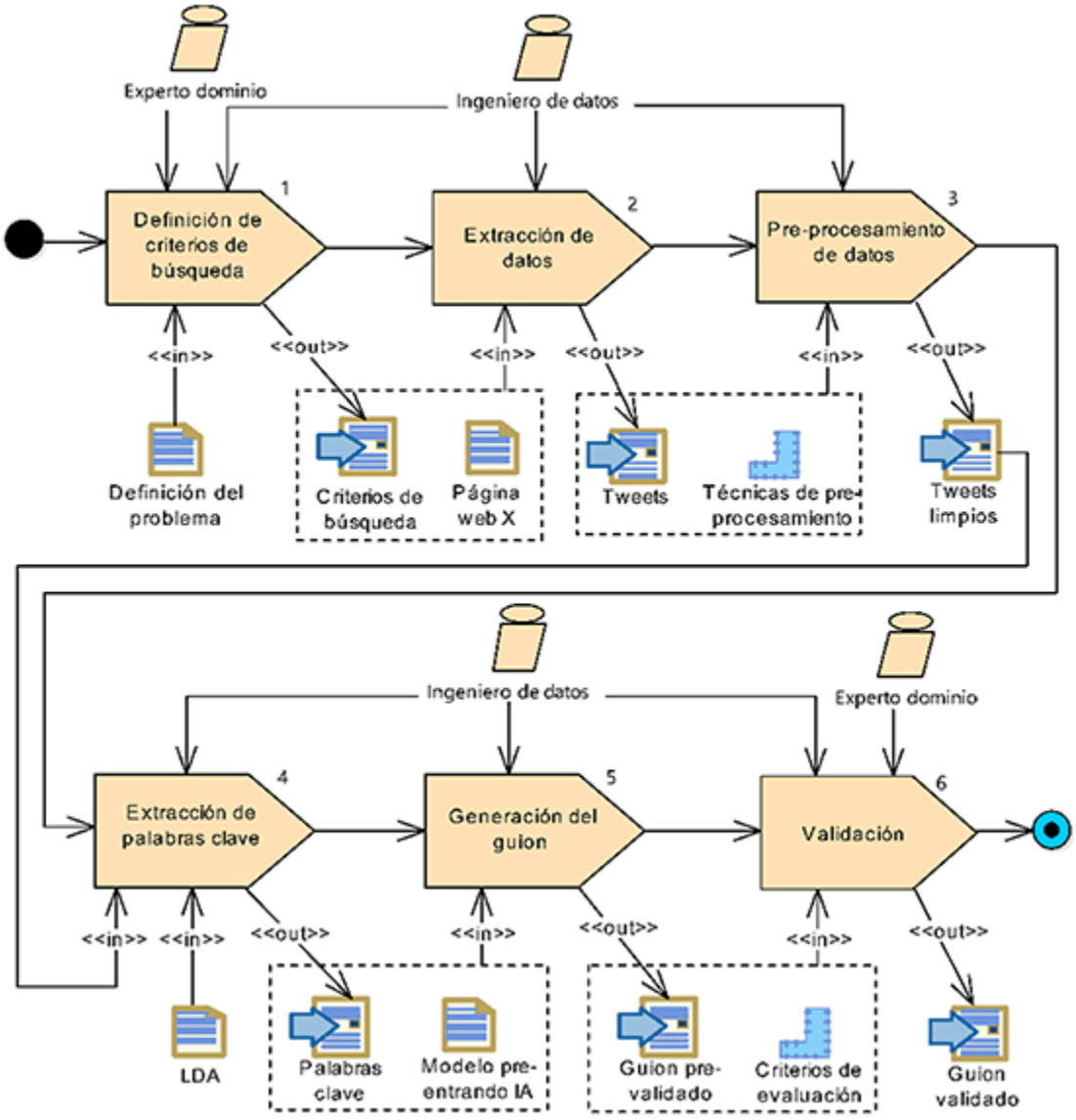

Figura 2

Diagrama SPEM de la metodología basada en CRISP-DM

Definición de criterios de búsquedaDefinición de criterios de búsqueda

En esta fase fueron establecidos los criterios y parámetros de búsqueda en la plataforma X para asegurar que los datos recolectados sean relevantes para el estudio. Estos criterios proporcionan la información necesaria para la creación de cadenas de búsqueda específicas de textos en la red social X y son fundamentales para obtener los datos requeridos. Fue necesario ajustar estos criterios a las capacidades de la opción de búsqueda avanzada de la plataforma. Entre los criterios que han sido definidos con los expertos en el

dominio de acoso escolar y ciberacoso están el tipo de publicación, palabras relevantes, hashtags, lenguaje y sección.

Extracción de datos

El objetivo de esta fase es la extracción de tweets desde la plataforma X utilizando los criterios de búsqueda definidos en la fase anterior. Esta fase es crucial para recopilar datos relevantes a ser analizados y utilizados en etapas posteriores del estudio. El proceso de búsqueda y extracción de datos de la plataforma fue automatizado a través de una interfaz gráfica desarrollada en el lenguaje de programación *Python* y *Selenium Web Driver*. Esta interfaz admite el ingreso de un texto de búsqueda, la descripción, la cantidad máxima de tweets a recuperar y las opciones para buscar en tweets recientes e incluir las respuestas.

Adicionalmente, para explorar los resultados de búsqueda en la página de X, se utilizó la función de desplazamiento continuo. Esto significa que, al llegar al final de la página, se cargarán automáticamente más post si están disponibles. Cuando no haya más resultados, el desplazamiento se detendrá en la parte inferior de la página. Internamente, la interfaz automatiza los siguientes procesos: i) validación de sesión, ii) petición web, iii) inclusión de respuestas y iv) visualización de resultados. A continuación, se describen los pasos de la automatización.

- Validación de sesión: Se accede a la página de inicio de X a través de su URL. Si la solicitud es exitosa, se confirma la sesión activa y se procede al siguiente paso. Caso contrario, si la sesión no ha sido iniciada, la solicitud redirigirá automáticamente a la página de inicio de sesión. De ser este el caso, automáticamente se rellenan los campos de credenciales y la cookie de sesión es almacenada en un archivo para futuros eventos.

- Petición web: Una solicitud web es generada con los parámetros y criterios definidos. Una vez cargados los tweets, se procedió a extraer y almacenar la información relevante. Para ello, ha sido aprovechada la estructura de la página web de X, donde cada tweet se encuentra en un elemento HTML con el atributo data-testid = “tweet”. Todos los elementos similares fueron identificados y se obtuvieron los detalles de cada tweet, como el texto, la cantidad de “me gusta”, re-tweets, enlaces, etc. Además, la función de desplazamiento fue utilizada para cargar los resultados hasta alcanzar la cantidad requerida o alcanzar el número de tweets existentes.

- Inclusión de respuestas: En caso de haber seleccionado la opción de incluir respuestas, a través de una petición web, se extraen las respuestas de cada tweet hasta alcanzar el número máximo de respuestas definido. Son almacenadas únicamente las respuestas al tweet original realizado por el mismo usuario, evitando así las respuestas irrelevantes o inapropiadas que no aportan al desarrollo del hilo.

- Visualización de resultados: Los resultados obtenidos en la búsqueda son visualizados en la interfaz gráfica.

Pre-procesamiento de datos

En esta etapa, tras finalizar la etapa de búsqueda y extracción, los datos fueron almacenados en un conjunto de datos y posteriormente en un objeto *DataFrame* de la librería *Pandas* para su manipulación mediante *Python*. Luego, se inició el pre-procesamiento de los datos recopilados con el objetivo de limpiar y preparar los textos con el objetivo de optimizar los resultados en la generación del guion. Para ello, en la Tabla 1 se describen las técnicas de PLN utilizadas.

Tabla 1

Técnicas de pre-procesamiento utilizadas

TÉCNICA	DEFINICIÓN
Convertir el texto a minúsculas.	Esta técnica transforma todas las letras del texto a minúsculas con el objetivo de normalizar y facilitar el análisis textual.
Eliminar hashtags.	Consiste en eliminar las palabras o frases precedidas con el símbolo "#" o hashtag. Si bien se utilizan para agrupar publicaciones, no tienen relevancia para el resto del estudio.
Eliminar correos electrónicos.	Esta técnica consiste en eliminar los correos electrónicos presentes en los registros, ya que no contribuyen al análisis posterior.
Eliminar signos de puntuación y caracteres especiales.	Tiene como objetivo reducir la dimensionalidad, simplificar y normalizar los datos. Consiste en eliminar de los registros todos los signos de puntuación como puntos, comas, signos de interrogación, etc.
Eliminar números.	Los números presentes en los registros no aportan ninguna información valiosa para el estudio, por lo tanto, con esta técnica, fueron eliminados todos los caracteres numéricos de los registros.
Eliminar Stopwords.	Consiste en remover palabras comunes de un texto que no aportan un significado relevante al estudio, como "el", "la", "de", etc. Esto reduce la dimensionalidad de los datos y destaca las palabras relevantes.
Lematizar.	Se reducen las palabras a su forma base o lema, que es la forma en la que se encuentra en un diccionario. Por ejemplo, el lema de "corriendo" es "correr".
Análisis de sentimientos.	Esta técnica tiene como objetivo determinar la actitud o emoción que expresa un texto. Esta actitud o emoción puede ser positiva, neutral o negativa, dependiendo de la herramienta utilizada.

Extracción de palabras claveExtracción de palabras clave

En esta etapa, se utilizó el modelo *LDA* de la librería *Gensim* en Python para identificar las palabras más relevantes en los textos. Para determinar el número óptimo de tópicos ha sido utilizado el *indicador de coherencia LDA*, buscando el valor más cercano a 1. Las pruebas fueron realizadas con un rango de 1 a 30 tópicos. Una vez identificado el número óptimo de tópicos con el mayor indicador de coherencia, se utilizaron las palabras más importantes de cada tópico.

Para este estudio, se ha establecido un límite de 50 palabras clave en total. Por ejemplo, si el número óptimo de tópicos fuera 10, serían seleccionadas 5 palabras de cada tópico.

Generación del guion

Para la generación del guion, fue utilizado el modelo generativo pre-entrenado Transformer de tercera generación denominado "text-davinci-003" de la empresa *OpenAI* a través de su API y el lenguaje de programación *Python*. Este modelo es ideal para tareas que requieren seguir instrucciones detalladas sin necesidad de ejemplos previos, además de manejar una amplia ventana de contexto de hasta 4,097 tokens. Su entrenamiento previo garantiza la generación de contenido relevante y coherente.

OpenAI ofrece recomendaciones para el uso efectivo de sus modelos de generación de texto. Para esta tarea, fueron consideradas las siguientes:

- Utilizar un modelo adecuado: "text-davinci-003" es recomendado para generación de texto y modelos como "code-davinci-002" para generación de código.
- Poner instrucciones al inicio: Separar las instrucciones del contexto mejora la efectividad del modelo.
- Ser específico: Detallar el contexto, salidas, longitud, formato, estilo, etc., para obtener resultados más precisos.

Por lo tanto, tomando en consideración las recomendaciones de la empresa en cuanto al uso efectivo de sus modelos de generación de texto, fueron identificados y establecidos los siguientes parámetros presentados en la Tabla 2.

Tabla 2

Parámetros para la generación del guion

PARAMETRO	DEFINICIÓN
Modelo	El modelo seleccionado, en este caso "text-davinci-003", la elección del modelo afecta el contenido generado y los costos de su utilización.
Prompt	Las instrucciones que ejecuta el modelo.
Temperatura	Un valor cercano a 1 genera respuestas más creativas, pero menos predecibles. Se utilizó un valor de 0.
Max tokens	Límite máximo de tokens permitidos, 3,000 tokens en este caso.
Top p	Valor máximo para la distribución de probabilidad de los tokens a usar. Se estableció un valor de 1.
Frecuency penalty	Penaliza tokens repetidos, se estableció un valor de 0.

Validación

Los resultados obtenidos fueron evaluados en colaboración con un experto en la materia para determinar si se cumplieron los objetivos planteados. La herramienta presenta los datos clave del análisis exploratorio realizado, así como el guion generado para la cápsula educativa. Para evaluar los resultados, se utilizó el enfoque de Meta-Pregunta-Respuesta (GQM, por sus siglas en inglés) (Van Solingen et al., 2002). La Tabla 3 presenta la meta de la evaluación y la Tabla 4 expone las preguntas formuladas al experto en la materia utilizando el enfoque antes mencionado. En caso de responder afirmativamente las preguntas, se considera el modelo como válido.

Tabla 3

Meta de la evaluación con el enfoque GQM

META	COMPONENTE
Evaluar	Modelo de generación de guion para cápsulas educativas.
Con el propósito de	Evaluar la calidad del guion generado.
Desde el punto de vista de	Experto en la materia.

Tabla 4

Preguntas y métricas de la evaluación con el enfoque GQM

PREGUNTA	MÉTRICAS
¿El guion generado cumple con sus expectativas en el contexto solicitado?	- Promueve el fortalecimiento de la autoestima. - Promueve la empatía.
¿Cumple las características de una cápsula educativa?	- Es un contenido que se puede dar en unidades pequeñas de tiempo. - Es fácil de comprender.
¿Usa las palabras solicitadas?	- El número de palabras solicitadas. - El número de palabras utilizadas.

Resultados y Discusión

Para una mejor comprensión, los resultados fueron divididos en cuatro etapas: i) definición de criterios de búsqueda, donde se exponen los criterios definidos luego del análisis con el experto en la materia; ii) extracción y pre-procesamiento de datos, donde se exponen los resultados de la extracción de datos y la aplicación de las técnicas de pre-procesamiento; iii) extracción de palabras clave, donde se exponen los resultados de la extracción de palabras clave con la técnica LDA y iv) generación y validación del guion, donde se expone el guion generado con el modelo pre-entrenado y su validación.

Definición de criterios de búsqueda

Los criterios de búsqueda fueron establecidos en colaboración con un experto en la materia de acoso escolar y ciberacoso, seleccionando información relevante para el estudio sobre el acoso escolar y el ciberacoso en la plataforma X. Se decidió excluir cualquier tipo de enlace de la búsqueda, ya que los usuarios suelen compartir contenido externo que no es relevante para este estudio. La Tabla 5 detalla los criterios definidos para la búsqueda.

Tabla 5

Resultados de los criterios de búsqueda

CRITERIO	DEFINICIÓN	CRITERIOS DEFINIDOS
Tipo de publicación	El tipo de publicación en la red social X: Tweets, Retweets, Hilos, Respuestas, etc.	Hilos.
Palabras relevantes	El contenido de las publicaciones a extraer.	“Bullying”, “Cyberbullying”, “Acoso escolar”, “Ciberacoso”.
Hashtags	Palabras o frases precedidas por el símbolo “#” utilizadas para categorizar y agrupar tweets.	#bullying, #cyberbullying, #acosoescolar, #ciberacoso, #ciberbullying, #acoso.
Lenguaje	Lenguaje de las publicaciones.	Español.
Sección	Nombre de la sección de la plataforma.	Secciones “Destacado” y “Más reciente”.

Adicionalmente, se decidió excluir las respuestas de la búsqueda, ya que el enfoque se centra en los hilos de conversación, que son respuestas a un tweet inicial. De esta manera, los textos de búsqueda que fueron utilizados son los siguientes:

- “hilo (#bullying) Lang:es -filter:links -filter:replies”
- “hilo (#cyberbullying) Lang:es -filter:links -filter:replies”
- “hilo (#acosoescolar) Lang:es -filter:links -filter:replies”
- “hilo (#ciberacoso) Lang:es -filter:links -filter:replies”
- “hilo (#acoso) Lang:es -filter:links -filter:replies”
- “hilo (#ciberbullying) Lang:es -filter:links -filter:replies”

Extracción y pre-procesamiento de datos

Durante el proceso de automatización de la búsqueda de datos, se filtraron varios anuncios entre los resultados; por ende, la cantidad de caracteres permitidos en un tweet aumentó de 280 a 4,000 caracteres. Debido a esto, se realizaron ajustes durante el proceso de extracción para mitigar estos problemas. De este modo, 1,645 tweets fueron recuperados. En la Tabla 6 se aprecia el número de tweets recuperados de acuerdo con los hashtags utilizados.

Tabla 6

Cantidad de tweets por hashtag

HASHTAG	CANTIDAD DE TWEETS
#bullying	921
#acosoescolar	342
#acoso	237
#ciberacoso	97
#cyberbullying	32
#ciberbullying	16

Tras el pre-procesamiento de los datos, el análisis de sentimientos reveló que 146 tweets fueron clasificados como positivos y 1,325 fueron clasificados como negativos. Esto indica que el tema del acoso escolar y ciberacoso se asocia predominantemente con sentimientos negativos en la plataforma X. Adicionalmente, la Figura 3 ilustra los hashtags más utilizados en los resultados de la búsqueda, y la Figura 4 ilustra las palabras más comunes en los resultados de la búsqueda.

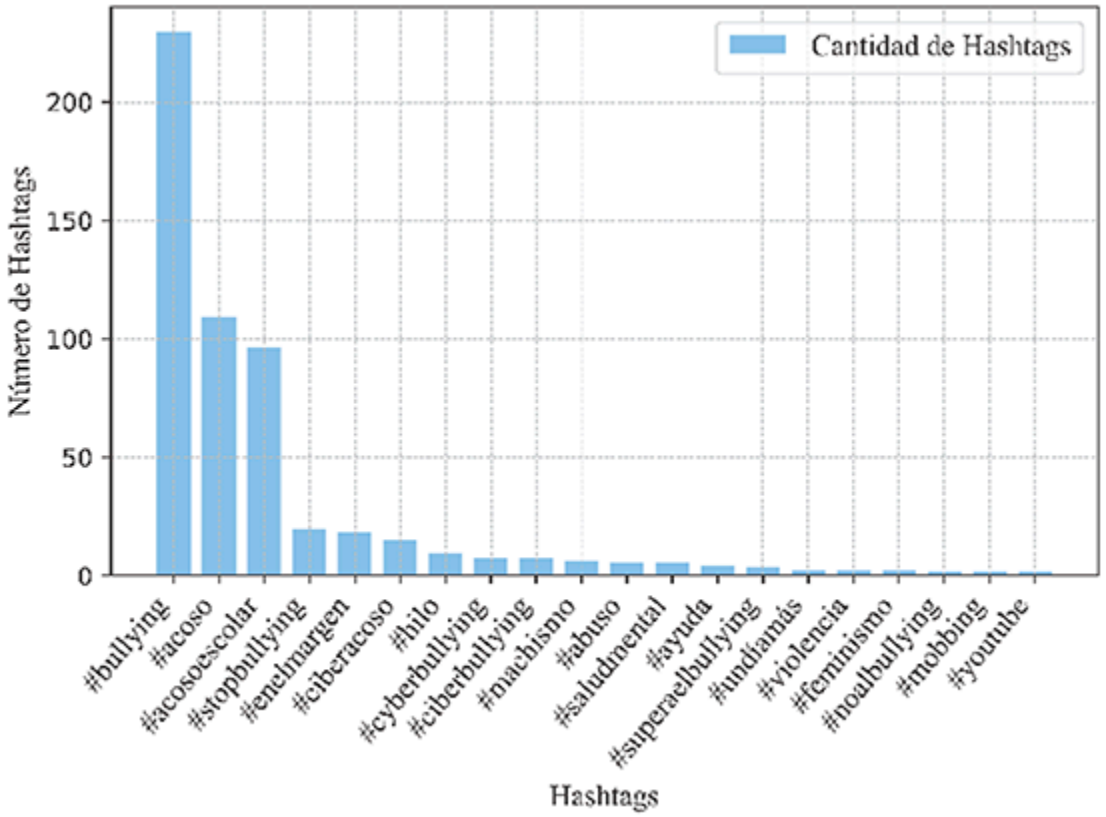

Figura 3

Hashtags más utilizados en los resultados de la búsqueda

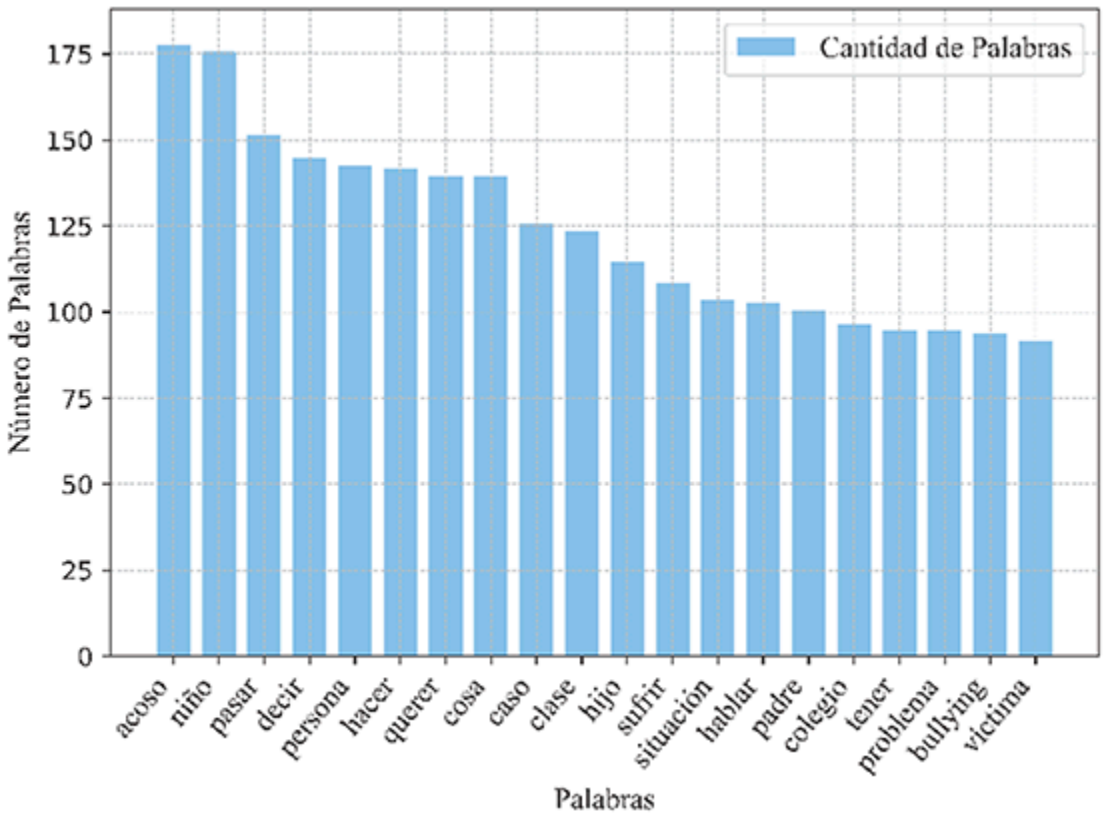

Figura 4

Palabras más populares en los resultados de la búsqueda

Es importante subrayar que el acceso a los datos de la plataforma X se ha convertido en un proceso complejo y costoso. Inicialmente, la API ofrecía un acceso sencillo y gratuito, pero los cambios recientes han limitado su uso. Esto ha impulsado a los usuarios a recurrir al web scraping como alternativa para obtener datos. Durante un tiempo, esta alternativa se convirtió en la principal fuente para obtener datos de X. Sin embargo, la plataforma identificó estos puntos de acceso y tomó medidas para restringirlos. Este cambio en la política de acceso a datos plantea nuevos desafíos para las investigaciones que dependen de la información de X.

Extracción de palabras clave

Utilizando la técnica LDA y un límite de 50 palabras clave, se calculó el indicador de coherencia para determinar el número óptimo de tópicos. La Figura 5 muestra los indicadores en un rango de 1 a 30 tópicos, donde un valor cercano a 1 indica la cantidad óptima de tópicos.

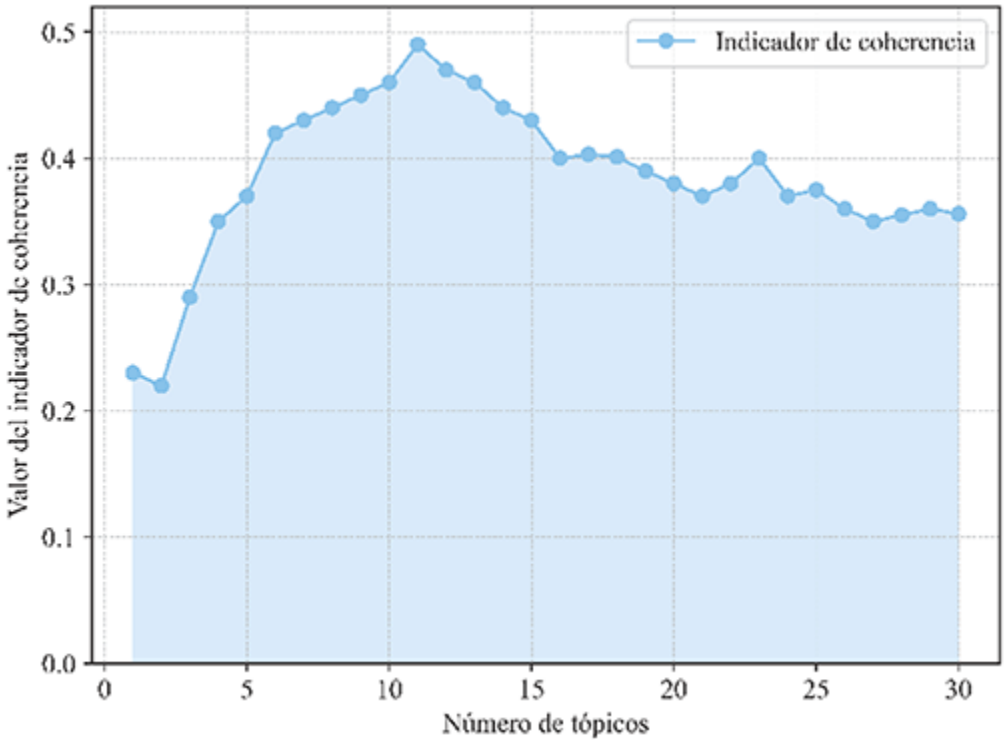

Figura 5

Indicador de coherencia para el número óptimo de tópicos

Como se observa en la Figura 5, el indicador de coherencia alcanza su punto máximo alrededor del número 11, lo que sugiere que el número óptimo de temas a considerar es 11.

Generación y validación del guion

Con el modelo seleccionado "text-davinci-003" de la empresa OpenAI se ha generado el guion para la cápsula educativa. Este modelo responde a instrucciones específicas para la creación de contenido que cumpla con los requisitos solicitados. Adicionalmente, el modelo sugirió recursos visuales que acompañan al texto del guion; estos elementos se presentan en la Tabla 7.

Tabla 7

Guion generado con el modelo "text-davinci-003"

NARRADOR	RECURSOS
Esta es la historia de un niño de 8 años.	(Imagen de un niño de 8 años)
Él vive con su familia en una casa normal.	(Imagen de una familia feliz)
Va a la escuela y trabaja duro para aportar algo a su comunidad.	(Imagen de un niño en la escuela)
Un día, él publicó una foto en una red social.	(Imagen de un niño en una computadora)
Pero alguien comienza a acosarlo.	(Imagen de un niño con una mirada triste)
El niño se siente triste y confundido.	(Imagen de un niño llorando)
Él decide hablar con alguien de su familia para pedir ayuda.	(Imagen de un niño hablando con un adulto)
Él denuncia el caso y lleva pruebas para respaldar su denuncia.	(Imagen de un niño hablando con un policía)
Él habla con sus amigos sobre el tema y les explica lo que deberían hacer si alguna vez se enfrentan a una situación similar.	(Imagen de un niño hablando con un grupo de amigos)
El niño se siente mejor al saber que hay gente que lo entiende y que puede ayudarlo.	(Imagen de un niño sonriendo)
Él aprende que siempre puede contar con su familia y amigos para seguir adelante.	(Imagen de un niño abrazando a un adulto)
Él aprende que la autoestima es importante y que siempre hay alguien que puede ayudar.	(Imagen de un niño sonriendo)

Gracias por trabajar juntos para entender este tema y ayudar a los niños a pensar de forma positiva.	(Música de fondo)

A simple vista, el guion generado cumple con las características esenciales de las cápsulas educativas: es conciso, claro y enfatiza los aspectos clave del tema, alineándose con el contexto requerido. Además, el experto en la materia ha respondido afirmativamente todas las peguntas de la evaluación preparada con el enfoque GQM, lo cual valida que el mensaje transmitido es adecuado para el objetivo de contribuir al aprendizaje y la prevención del ciberacoso con el formato de cápsulas educativas, tal como se muestra en la Tabla 8.

Tabla 8

Evaluación de resultados con el enfoque GQM

PREGUNTA	MÉTRICAS	RESULTADOS
¿El guion generado cumple con sus expectativas en el contexto solicitado?	- Promueve el fortalecimiento de la autoestima.	- Positivo.
	- Promueve la empatía.	- Positivo.
¿Cumple las características de una cápsula educativa?	- Es un contenido que se puede dar en unidades pequeñas de tiempo.	- Positivo.
	- Es fácil de comprender.	- Positivo.
¿Usa las palabras solicitadas?	- El número de palabras solicitadas.	- 27 palabras de 36.
	- El número de palabras utilizadas.	- Usa el 75% de las palabras.

Conclusiones

A pesar de las limitaciones para la extracción de datos, este estudio ha demostrado la factibilidad de utilizar modelos pre-entrenados en la creación de cápsulas educativas para el combate del ciberacoso. La técnica LDA resultó efectiva en la extracción de palabras clave de textos cortos, siempre y cuando se realice un adecuado pre-procesamiento de los datos. En cuanto a la generación de contenido textual, se demostró la eficacia del modelo “text-davinci-003” en la creación de guiones coherentes y relevantes a partir de instrucciones precisas.

Adicionalmente, el proceso de extracción de datos relacionados con el acoso escolar y el ciberacoso ha sido automatizado. Se destacó la

importancia de definir adecuadamente los términos de búsqueda para obtener datos relevantes, considerando la evolución y actualización continua de la plataforma. A través de un análisis riguroso y la colaboración con un experto en la materia, se validaron los resultados obtenidos, confirmando que las cápsulas educativas generadas transmiten un mensaje adecuado para prevenir y combatir el ciberacoso.

Este estudio demuestra el potencial de combinar técnicas de minería de texto y modelos de lenguaje pre-entrenados para generar contenido educativo relevante y efectivo en la lucha contra el ciberacoso. Sin embargo, es necesario considerar los costos tanto de la extracción de datos como del uso del modelo pre-entrenado. Actualmente, existen varios modelos pre-entrenados que no requieren suscripción, son de acceso gratuito y pueden ser utilizados para propósitos similares. Igualmente, es posible incorporar nuevas técnicas de extracción de palabras clave y comparar su efectividad. De esta manera, las cápsulas educativas pueden ser generadas de una forma más económica y en mayor cantidad, contribuyendo activamente a la lucha contra el ciberacoso.

A futuro, se propone ampliar el alcance de este estudio explorando alternativas más accesibles y diversas. En particular, se plantea evaluar el potencial de modelos de lenguaje libres para la generación de guiones de cápsulas educativas. Este tipo de modelos libres podrían representar una solución viable en términos de recursos económicos, democratizando la creación de contenido educativo. Además, se propone la experimentación de otras técnicas de extracción de palabras clave o modelos pre-entrenados basados en la arquitectura *Transformers*. La comparación del rendimiento de estas técnicas con la técnica *LDA* permitirá determinar la más adecuada para el objetivo de la investigación. Finalmente, se propone la exploración de diversas fuentes de datos, como otras plataformas de redes sociales, donde se podría extraer contenido relacionado con el acoso escolar y el ciberacoso.

Reconocimientos

Los autores desean agradecer al Vicerrectorado de Investigaciones de la Universidad del Azuay por el apoyo financiero y académico, así como a todo el personal de la escuela de Ingeniería de Ciencias de la Computación y el Laboratorio de Investigación y Desarrollo en Informática (LIDI).

Referencias

Alim, S. (2015). Analysis of tweets related to cyberbullying: Exploring information diffusion and advice available for cyberbullying victims. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 5*(4), 31–52.

Arazzi, M., Nicolazzo, S., Nocera, A., & Zippo, M. (2023). The importance of the language for the evolution of online communities: An analysis based on Twitter and Reddit. *Expert Systems with Applications, 222*, 119847.

Azuela, J. H. S., & Ayala, A. P. (2019). *ESTADO DEL ARTE EN INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS*.

Bayari, R., & Bensefia, A. (2021). Text mining techniques for cyberbullying detection: state of the art. *Adv. sci. technol. eng. syst. j, 6*(1), 783–790.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.

Chen, K., Duan, Z., & Yang, S. (2022). Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences, 41*(1), 114–130.

Fatima, N., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Soomro, A. (2022). A systematic literature review on text generation using deep neural network models. *IEEE Access, 10*, 53490–53503.

Garaigordobil, M. (2014). Cyberbullying. Screening de acoso entre iguales: descripción y datos psicométricos. *International Journal ot Developmental and Educational Psychology, 4*(1), 311–318.

Guallar, J., & Traver, P. (2020). Curación de contenidos en hilos de Twitter. Taxonomía y ejemplos. *Anuario ThinkEPI, 14*.

Kamilali, D., & Sofianopoulou, C. (2015). Microlearning as Innovative Pedagogy for Mobile Learning in MOOCs. *International Association for Development of the Information Society*.

Lugones Botell, M., & Ramírez Bermúdez, M. (2017). Bullying: aspectos históricos, culturales y sus consecuencias para la salud. *Revista cubana de medicina general integral, 33*(1), 154–162.

Mancilla-Vela, G., Leal-Gatica, P., Sanchez Ortiz, A., & Vidal, C. (2020). Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: un análisis en minería de datos. *Formación*

universitaria, 13, 23–36. https://doi.org/10.4067/S0718-50062020000600023

Orellana, M., Zambrano-Martinez, J. L., Calle Andrade, R. M., Roldan, A., & Tirado Jarama, A. N. (2023). Generación de Texto Guía para la Detección Automatizada del Acoso y el Ciberacoso. *Revista Tecnológica - ESPOL, 35*(2), 181–191. https://doi.org/10.37815/rte.v35n2.1049

Peña, A., & Herrera, L. (2021). Indicadores de tecnología de la información y comunicación. *Quito: INEC*.

Salmivalli, C., Laninga-Wijnen, L., Malamut, S. T., & Garandeau, C. F. (2021). Bullying prevention in adolescence: Solutions and new challenges from the past decade. *Journal of Research on Adolescence, 31*(4), 1023–1046.

Sanchez, H., & Kumar, S. (2011). Twitter bullying detection. *ser. NSDI, 12*(2011), 15.

Van Solingen, R., Basili, V., Caldiera, G., & Rombach, H. D. (2002). Goal question metric (gqm) approach. *Encyclopedia of software engineering*.

Vázquez, A., Pinto, D., Vilariño, D., & Castro, M. (2017). Modelos para la generación automática de diálogos: Una revisión. *Applications of Language & Knowledge Engineering*, 163.

Vidal Ledo, M., Vialart Vidal, M. N., Alfonso Sánchez, I., & Zacca González, G. (2019). Cápsulas educativas o informativas. Un mejor aprendizaje significativo. *Educación médica superior, 33*(2).

William Hermel Astudillo Quituisaca, Priscila Cedillo, Marcos Orellana

Extracción de Palabras Clave de Ciberacoso de Textos Breves: un Enfoque de Aprendizaje Automático

Extracting Cyberbullying Keywords from Short Texts: A Machine Learning Approach

Revista Tecnológica ESPOL - RTE
vol. 36, núm. 1, Esp. p. 25 - 38, 2024
Escuela Superior Politécnica del Litoral, Ecuador
rte@espol.edu.ec

ISSN: 0257-1749
ISSN-E: 1390-3659

DOI: https://doi.org/10.37815/rte.v36nE1.1207