UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO
MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

DOCTORADO EN CIENCIAS

MODELO DE CLASIFICACIÓN REPRESENTATIVO DE SITUACIONES DE CYBERBULLYING

TESIS PROFESIONAL PARA CUBRIR PARCIALMENTE LOS REQUISITOS PARA OBTENER EL GRADO DE DOCTOR EN CIENCIAS

PRESENTA:
KARLA IVETTE ARCE RUELAS

DIRECTOR DE TESIS:
DRA. LILIANA CARDOZA AVENDAÑO

CO-DIRECTOR DE TESIS:
DR. OMAR ÁLVAREZ XOCHIHUA

ENSENADA, BAJA CALIFORNIA, MÉXICO, ABRIL DEL 2023

Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Modelo de clasificación representativo de situaciones de cyberbullying

TESIS PROFESIONAL

que para cubrir parcialmente los requisitos para obtener el grado de

DOCTOR EN CIENCIAS

Presenta

Karla Ivette Arce Ruelas

Dra. Liliana Cardoza Avendaño
Director de Tesis

Dr. Omar Álvarez Xochihua
Co-Director de Tesis

Dr. José Ángel González Fraga
Miembro del Comité

Dra. Rosa Martha López Gutiérrez
Miembro del Comité

Dr. César Cruz Hernández
Miembro del Comité

Ensenada, Baja California, México, Abril del 2023

RESUMEN de la Tesis de **Karla Ivette Arce Ruelas**, presentada como requisito parcial para la obtención del grado de DOCTOR EN CIENCIAS del programa Maestría y Doctorado en Ciencias e Ingeniería (MYDCI), de la Universidad Autónoma de Baja California. Ensenada, Baja California, México. Abril del 2023.

Modelo de clasificación representativo de situaciones de cyberbullying

Resumen aprobado por:

Dra. Liliana Cardoza Avendaño
Director de Tesis

Dr. Omar Álvarez Xochihua
Co-Director de Tesis

El acoso u hostigamiento, también referido internacionalmente como bullying, es una práctica que se ha presentado desde los inicios de las relaciones interpersonales. Sin embargo, en la actualidad, han incrementado los casos registrados de abusos y agresiones, esto debido a una nueva modalidad de bullying que se presenta mediante el uso de medios electrónicos, conocida como cyberbullying. Al mismo tiempo, una variedad de herramientas tecnológicas ha sido utilizada para la detección y prevención de este fenómeno social. En esta investigación se presenta un análisis del trabajo reportado en la literatura sobre las tecnologías utilizadas actualmente para la detección y atención de esta problemática. Siendo este problema un fenómeno estudiado y discutido a nivel nacional e internacional que aún mantiene abiertas diversas líneas de investigación. En las investigaciones actuales se enfatiza en la detección de comentarios específicos con contenido agresivo (agressive posts) con base en el análisis de su contenido. No considerando, en el análisis de cyberbullying, un enfoque de análisis multivariado, como la interrelación de comentarios agresivos con otros factores que influyen en la presencia de acoso cibernético, así como la relevancia de la fuente y el contenido del conjunto de datos utilizado para generar modelos que realmente representen entornos virtuales de interacción social entre jóvenes. En la presente investigación se describe el proceso de creación de un conjunto de datos representativo de entornos reales de interacción social en redes sociales, y se enfatiza la importancia de la fuente de datos. Posteriormente, se presentan las herramientas y algoritmos, propios de las áreas de aprendizaje automático y aprendizaje profundo, utilizados para el desarrollo de un modelo predictivo para detectar situaciones de cyberbullying considerando el contenido de las conversaciones y factores externos que influyen en su materialización. Los resultados obtenidos muestran que, ante un conjunto de datos reducido, un modelo convencional de aprendizaje automático arroja mejores resultados que modelos más complejos generados mediante técnicas de aprendizaje profundo.

Palabras clave: Bullying, cyberbullying, aprendizaje automático, redes sociales, aprendizaje profundo.

ABSTRACT of the thesis, presented by **Karla Ivette Arce Ruelas**, in order to obtain the DOCTORAL in SCIENCE DEGREE of the MYDCI program in Universidad Autónoma de Baja California. Ensenada, Baja California, México. Abril del 2023.

Representative classification model of cyberbullying situations

Approved by:

Dra. Liliana Cardoza Avendaño
Director de Tesis

Dr. Omar Álvarez Xochihua
Co-Director de Tesis

Bullying is a practice that has been present since the beginning of interpersonal relationships. However, currently registered cases of abuse and aggression have increased due to a new form of bullying that occurs through electronic media, known as cyberbullying. At the same time, various technological tools have been used to detect and prevent this social phenomenon. This research analyzes the work reported in the literature on the technologies currently used for the detection of this problem. A phenomenon studied and discussed at a national and international level that still keeps various research lines open, highlighting how current research emphasizes the detection of specific publications with aggressive content (aggressive posts). Not considering, in the analysis of cyberbullying, aspects such as the interrelationship of the set of cyberbullying actions, situations that trigger it, and the relevance of the source and content of the dataset used to generate what really represent virtual environments of social interaction among young people. This research describes the process of creating a representative dataset of natural and real environments of social interaction in social networks. Subsequently, typical tools and algorithms of machine learning and deep learning are used to develop a predictive model to detect potential cyberbullying scenarios before their materialization. As a result, a conventional machine learning model reported better results than more complex models generated by deep learning techniques.

Keywords: Bullying, cyberbullying, machine learning, social networks, deep learning.

Dedicatoria

Quiero dedicar este trabajo:

A Dios, quién guía mi camino y me da la fortaleza para continuar.

A mis padres, por su amor y compresión, por que con gran esfuerzo me dieron la oportunidad de llegar hasta donde estoy.

A mi querido hermano por ser mi apoyo incondicional.

A Lucas por ser mi más fiel compañero de desvelos y nunca dejarme sola.

A Balú por recordarme sonreír aún en momentos de estrés.

A mi Esposo, por sus palabras de apoyo y confianza. Por creer en mi y en cada uno de mis sueños. Por su amor y por ser mi compañero de nuevas aventuras.

A todos ustedes, con amor.

Karla Ivette Arce Ruelas.

Agradecimientos

Quiero expresar mi agradecimiento:

A mis sinodales Dr. Omar Álvarez Xochihua, Dra. Liliana Cardoza Avendaño, Dr. José Ángel González Fraga, Dra. Rosa Martha López Gutiérrez y Dr. César Cruz Hernández, por sus consejos, críticas y apoyo durante el desarrollo de esta tesis.

Al personal administrativo de la Facultad de Ingeniería, Arquitectura y Diseño y el personal del programa de Maestría y Doctorado en Ciencias e Ingeniería (MyDCI) por su atención, amabilidad y disposición para ayudarme.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico brindado mediante la beca doctoral con número de becario 588893 y la Beca de Investigación 166654 (A1-S-31628).

A la Facultad de Ingeniería Arquitectura y Diseño de la Universidad Autónoma de Baja California (UABC) por las facilidades otorgadas para la realización de este trabajo.

ÍNDICE

Tabla de contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

A lo largo de la humanidad, las relaciones interpersonales se han considerado como asociaciones complejas. Entre ellas destacan comportamientos afectivos, de indiferencia y agresivos, siendo la conducta agresiva un factor que ha generado repercusiones negativas en la sociedad, provocando situaciones de estrés, ansiedad o depresión (Loredo-Abdalá et al., 2008). Generalmente, un comportamiento agresivo puede llegar a manifestarse mediante acciones físicas o verbales, ya sea entre pares de individuos o en relaciones grupales. Presentándose esta situación, indistintamente, en diferentes ambientes sociales de interacción: escuela, espacios recreativos, trabajo, hogar, entre otros.

Particularmente en el ámbito educativo, estas interacciones agresivas se categorizan como un tipo de acoso psicológico o moral que es conocido internacionalmente con el término en inglés *bullying* (acoso), término utilizado en el resto de la tesis para referir a este problema social. En la literatura existen diversas definiciones del concepto de *bullying*, todas ellas coinciden en 3 elementos: 1) que es toda aquella acción o comportamiento agresivo, 2) realizado de manera repetitiva, y 3) con la intención de dañar física o emocionalmente a una persona que se encuentre en desventaja (Continente et al, 2010).

La práctica de *bullying* se ha diversificado y ha adquirido nuevas dimensiones, esta problemática se ha trasladado y potenciado a entornos de comunicación basados en tecnologías de la información y la comunicación (TIC); principalmente mediante el uso de Internet y las redes sociales. En otras palabras, el avance tecnológico ha incrementado el alcance y las categorías de la práctica de *bullying*. En el ámbito educativo, derivado del avance tecnológico, el acoso escolar ya no se limita a la agresión presencial realizada durante las horas de escuela, sino que también mediante una modalidad donde no existen fronteras que limiten su alcance. Se ha creado un escenario donde la víctima no tiene donde esconderse, y se encuentra expuesta ante una mayor cantidad de público de forma inmediata (Benavides, 2015). Permitiendo al acosador realizar *bullying*, de manera directa o anónima, mediante mensajes de texto, grabaciones de audio, imágenes ofensivas, videos, entre otros. Esta nueva forma de acoso recibe el nombre de *cyberbullying* (ciberacoso).

El *cyberbullying* tiene los mismos efectos que el *bullying* tradicional, daña la confianza y la autoestima de la víctima, provocando ansiedad, frustración e inclusive ideas suicidas. Ante esta variación y de acuerdo con múltiples estudios científicos en el ámbito internacional (Lessne & Yanez, 2016; Modecki et al., 2014), se pone en evidencia la importancia de atender este problema a través de la misma tecnología que lo genera. Se deben crear y adaptar desarrollos tecnológicos que permitan detectar, clasificar y, de ser posible, evitar escenarios de ciberacoso.

Con el objetivo principal de actuar en contra de este tipo de agresión y proveer atención inmediata a la víctima, investigaciones recientes han dado énfasis en el análisis de texto que proviene de foros virtuales de discusión donde interactúan diferentes grupos de

personas (Al-Garadi et al., 2019). Lo anterior, mediante técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés), algoritmos propios de aprendizaje automático (ML, por sus siglas en inglés) y aprendizaje profundo (DL, por sus siglas en inglés).

En ML un modelo es el resultado del aprendizaje de un algoritmo al analizar y ejecutar un conjunto de datos. Un modelo representa lo que aprendió un algoritmo de aprendizaje automático, así como las reglas, números o estructuras de datos específicas resultantes de ejecutar el algoritmo que permitirán realizar clasificaciones o predicciones de acuerdo con un nuevo conjunto de datos (Rojas, 2020).

En la actualidad se ha buscado generar modelos computacionales que permitan representar y estudiar el fenómeno del ciberacoso. En la literatura se reportan modelos para la detección de expresiones con contenido de cyberbullying con diferentes niveles de precisión, que van del rango de 0.45 hasta un 0.95 (siendo 0.45 una baja precisión en la clasificación y 0.95 alta precisión) al clasificar un comentario agresivo (Rosa et al., 2019). La amplia variación entre los resultados reportados deriva principalmente de diversos factores, tales como el proceso de etiquetado de los datos de entrenamiento, el algoritmo utilizado, el origen y las características del corpus de datos.

Muchos de estos estudios están orientados a detectar escenarios de ciberacoso entre jóvenes en entornos educativos de nivel medio superior y superior. Sin embargo, uno de los principales problemas detectado fue que la mayoría de las investigaciones utilizan corpus lingüísticos con contenido agresivo provenientes de fuentes de libre acceso, las cuales permiten obtener corpus de gran tamaño, pero suelen no ser totalmente representativas de la población objetivo. En particular, se ha detectado el uso de fuentes de datos de poco uso por los jóvenes que asisten a instituciones educativas. Así como, la omisión común del análisis de escritura pictográfica (ej. emoticonos), de conversaciones acompañadas de datos multimodales (audio, imágenes, memes y/o videos) y del análisis multivariado a la hora de generar modelos predictivos; para nuestro caso de estudio, modelos de clasificación binaria que determinen si un comentario es considerado agresivo o no.

En este trabajo se realizó la creación y análisis de un corpus de cyberbullying en idioma español mexicano, donde se tomó en cuenta la procedencia de la fuente de los datos (ambientes privados de interacción) y la representatividad de las conversaciones con presencia de acoso. Adicionalmente, se reporta el desempeño obtenido al implementar algoritmos de clasificación tradicionales de aprendizaje automático y aprendizaje profundo, útiles para la clasificación de cyberbullying en entornos reales. Considerando la desventaja de contar con un corpus de tamaño reducido, pero con calidad y representatividad en su contenido, se obtiene un desempeño aceptable comparado con lo reportado en la literatura.

1.1 Planteamiento del problema

El acoso escolar, mayormente conocido como bullying, se caracteriza por constantes agresiones físicas, verbales y conductuales realizadas de forma intencional por una o más personas hacia sus compañeros (con alta incidencia en entornos educativos de niveles superior

y medio superior), generando daños físicos o psicológicos (Olweus, 1997). En los últimos años, esta práctica se ha extendido a ambientes de socialización electrónica, mayormente conocida como cyberbullying (Continente et al, 2010).

Tanto en el entorno presencial como en el electrónico, en atención a este problema, se realizan actividades de categorización, prevención, detección oportuna y corrección; sin embargo, son mínimas o pocas las acciones efectivas que permiten detectar la aparición, magnitud y características de este. Específicamente, en el estudio de cyberbullying, múltiples publicaciones científicas en el ámbito internacional (Lessne & Yanez, 2016; Modecki et al., 2014), evidencian la importancia de atender este problema a través de la misma tecnología que lo genera. A este respecto, la mayoría de las investigaciones, mediante el uso de técnicas de inteligencia artificial (IA), en áreas como minería de datos y aprendizaje automático, han logrado definir *modelos de clasificación* que permiten identificar, con una confiabilidad de hasta el 90%, acciones de cyberbullying, pero dichos modelos son generalmente entrenados con datos poco representativos, que difícilmente pueden ser trasladados para su uso a un entorno real de interacción social. Existe una mínima atención en la generación de *modelos de clasificación* que utilicen un conjunto de datos realmente característico de la audiencia objetivo, representativo con respecto al contenido y la fuente de los datos que permita lograr una identificación confiable de situaciones de cyberbullying, ya que este conjunto de datos representativo debe estar constituido por el conjunto de comentarios y acciones utilizadas de manera repetitiva por jóvenes estudiantes en entornos de redes sociales. Este conjunto de acciones mencionadas comprende: 1) las peculiaridades del lenguaje utilizado por el usuario, 2) el tipo de imágenes compartidas y 3) el uso de otros complementos de comunicación, como los emoticonos que utilizan en sus conversaciones; así como las posibles combinaciones entre ellas. Siendo de principal interés en el presente trabajo doctoral, la generación de modelos y algoritmos que nos permitan identificar escenarios de cyberbullying utilizando datos y factores adicionales provenientes de fuentes que realmente representen a la audiencia objetivo.

1.1.1 Justificación

Aunado a las mejoras y aportaciones que el avance tecnológico ha traído a nuestras actividades diarias, ya sean educativas, laborales o recreativas, situaciones desfavorables se han hecho presentes o bien han evolucionado. Uno de los avances tecnológicos que se ha identificado como un punto de inflexión en la existencia del ser humano es el Internet, y específicamente la aparición de las redes sociales; tecnología que aún se encuentra en su fase inicial de evolución.

Los beneficios principales de las redes sociales, a lo que también se conoce como Web-Social, son amplios y diversos, entre ellos encontramos: la oportunidad de socializar síncrona y asíncronamente con amigos, familiares, compañeros, conocidos y hasta con desconocidos que se encuentran físicamente cercanos o lejanos. Así como, complementar actividades educativas y comerciales e influir en acciones políticas y sociales. Por otro lado, esta misma tecnología ha detonado una nueva forma de agresión social conocida como cyberbullying.

Con el mismo potencial que las redes sociales impulsan las acciones favorables de socialización, se ha visto el impacto negativo que el cyberbullying ocasiona en la comunidad; evolucionando de un bullying tradicional, donde la presencia física era primordial, a un entorno donde no existe barrera física ni de tiempo. Al mismo tiempo, la diversidad y disponibilidad de múltiples medios (ej. videos, audios e imágenes) potencia la creatividad en las agresiones, y el suceso puede ser visto una y otra vez por agresores, agredidos y observadores (generando una sensación de escenificar una y otra vez esas situaciones desfavorables). Investigaciones actuales en el área de inteligencia artificial se han enfocado en la elaboración de modelos de clasificación para la detección y categorización de este fenómeno y sus participantes.

Al mismo tiempo, las interacciones humanas son complejas por sí solas, ya que influye una gran cantidad de variables determinadas por las características individuales de los participantes, e incluso algunas de estas variables son generadas de la misma interacción entre personas. El comportamiento del bullying y cyberbullying comprende una gran cantidad de variables que se presentan de acuerdo al contexto, como el empleo de datos multimedia, y los tipos de interacciones de los usuarios, por lo que podemos determinar que nos enfrentamos ante un escenario multifactorial. Aspecto poco estudiado en investigaciones actuales.

En este trabajo se complementan dichos estudios mediante la generación de *modelos de clasificación*, que permiten predecir escenarios de cyberbullying con un nivel de precisión superior al 80%, logrando identificar situaciones o escenarios de cyberbullying que es de utilidad para actuar previo al incremento de su intensidad; como bien menciona la definición de bullying, considerando que éste es una actividad agresiva que se realiza de manera repetitiva. Las situaciones o escenarios de cyberbullying comprenden precisamente dicha definición al referirse a cualquier actividad e interacción agresiva realizada de manera repetitiva transmitida mediante texto u otros materiales multimedia. Lo anterior, se aborda mediante la generación de modelos que hacen uso de algoritmos de aprendizaje automático y aprendizaje profundo que generan su base de conocimiento derivada de datos multimedia (ej. texto, emoticonos, y combinación de caracteres); obtenidos de espacios virtuales representativos, como entornos de redes sociales usados comúnmente por estudiantes de nivel medio superior y superior. Los resultados del presente estudio impactan en dos sentidos: 1) en el ámbito social, al disponer de un modelo, que puede ser trasladado a una aplicación social de usuario final, que permite identificar situaciones potenciales de cyberbullying previo el incremento de su intensidad, es decir, antes de afectar negativamente en miembros de su comunidad; y 2) en el área de estudio de IA, al construir y evaluar modelos y algoritmos para su generación y operación, derivados de datos en menor escala, pero provenientes de redes sociales representativas.

1.1.2 Objetivos de la investigación

1.1.2.1 Objetivo general

Crear un *modelo de clasificación*, robusto y escalable, es decir que pueda ser adaptado para incluir otros lenguajes o jergas y una mayor cantidad de datos a analizar, que permita

identificar situaciones de cyberbullying en espacios virtuales, construido y validado con datos multimedia que estos mismos entornos generan.

1.1.2.2 Objetivos específicos

1. Precisar el nivel de impacto de los factores que detonan los comportamientos de bullying y agresión cibernética en estudiantes de habla hispana del nivel educativo medio superior y superior, particularmente en la presencia de este fenómeno social en estudiantes de nuestro municipio, pudiendo extenderse a nivel estatal y nacional.
2. Determinar los patrones generados ante la combinación de acciones realizadas por el usuario, como el lenguaje utilizado y los emoticonos que utiliza en sus conversaciones, que desencadenen ***situaciones o escenarios*** potenciales de cyberbullying.
3. Identificar los espacios virtuales utilizados en mayor medida por la población objetivo, tales como Twitter, Facebook, Gmail y blogs.
4. Generar el *modelo de clasificación* que permita manipular datos multimedia en los espacios virtuales de mayor uso por la población objetivo, y obtener un nivel de precisión igual o mayor al 80%.
5. Diseñar los algoritmos que permitan implementar y manipular dicho modelo como una aplicación de usuario final.
6. Desarrollar una aplicación que permita monitorear las situaciones o escenarios potenciales de generar cyberbullying en el entorno deseado (ej. escuela, hogar, laboral).

1.1.3 Preguntas de investigación

Considerando la problemática mencionada, así como el avance tecnológico actual, las preguntas de investigación que sustentan nuestro estudio son:

1. ¿Qué factores detonan y permiten identificar situaciones de cyberbullying en entornos de redes sociales?
2. ¿Qué modelo de aprendizaje automático o aprendizaje profundo nos permite realizar una mejor clasificación de los factores detonantes de cyberbullying, en términos de precisión?

1.1.4 Hipótesis

Hipótesis principal:

Mediante el uso de técnicas de modelado actuales, en el área de inteligencia artificial, es posible identificar situaciones de cyberbullying con un nivel de precisión igual o mayor del

80%, impactando positivamente en los índices de detección y predicción de intensidad de este problema social en el nivel educativo medio-superior y superior.

Específicamente, suponemos que:

H1. Existen factores que gestan la materialización de cyberbullying en entornos de redes sociales derivados de datos multimedia, los cuales pueden ser identificables y medibles.

H2. Variantes de los modelos de aprendizaje automático y aprendizaje profundo, usados en inteligencia artificial, tal como el uso de variantes de redes neuronales, clusters, redes bayesianas, árboles de decisión u otra similar, permitirán predecir escenarios de cyberbullying en entornos de redes sociales con el nivel de confianza esperado.

1.2 Estructura de la tesis

Este trabajo está dividido en los siguientes capítulos:

Capítulo 1 En este capítulo se presenta una introducción del trabajo de tesis, se indica el planteamiento del problema, la justificación, el objetivo tanto general como específicos, preguntas e hipótesis de la investigación, y la secuencia del documento de trabajo de investigación.

Capítulo 2 Se presentan los antecedentes de investigaciones realizadas en el área de bullying y cyberbullying que contribuyen a sustentar los objetivos de este trabajo.

Capítulo 3 Aquí se detalla el enfoque metodológico en que se sustenta esta investigación basado en la metodología para la implementación de modelos de aprendizaje automático, la ***Machine Learning Pipeline (ML pipeline)***, que consiste en una serie de pasos ordenados que establecen la secuencia de trabajo de aprendizaje automático.

Capítulo 4 En este capítulo se presenta la relevancia de contar con un corpus de datos proveniente de redes sociales representativas de los sujetos de estudio objetivo. Se analiza la calidad de los datos y se eliminan datos atípicos. Finalmente como resultado de la metodología seguida y los datos obtenidos se construyó el conjunto final de archivos que conforman el corpus representativo con presencia de situaciones reales de cyberbullying. Subsecuentemente, en esta sección se describen las características del contenido del corpus lingüístico generado y se evalúa la representatividad para crear modelos de clasificación con desempeño aceptable.

Capítulo 5 En este capítulo se evalúa lel modelo de clasificación seleccionado que logró el mejor desempeño en un ambiente con usuarios reales controlado, se buscó corroborar su funcionamiento en una aplicación de usuario final, categorizando en tiempo real mensajes realizados en conversaciones expuestas a situaciones de cyberbullying.

Capítulo 6 Este capítulo finaliza con las conclusiones obtenidas durante el proceso del diseño, desarrollo y evaluaciones de la investigación, así como la respuesta a las preguntas de investigación y los objetivos que la rigen.

Capítulo 2

Antecedentes

Las agresiones remontan a los inicios de la existencia del ser humano, donde la supervivencia dependía directamente de la fuerza del hombre y de su capacidad para agredir o defenderse de sus contrincantes o depredadores. En aquellos tiempos los actos violentos y agresiones determinaban si un hombre era capaz de sobrevivir, siendo una actividad vital para defenderse ante situaciones de ataque, cazar o alimentarse. Sin embargo, al paso del tiempo, los actos de violencia ya no fueron principalmente con un sentido de sobrevivencia, pero siguieron siendo utilizados como acciones para aprovecharse, atemorizar o herir al débil. Arnold H. Buss (1961), definió las agresiones como una respuesta que provoca estímulos nocivos en otro organismo. Esta definición muestra la relación directa entre la acción del agresor y su objetivo, con la finalidad de dañar a su víctima. Adicionalmente, apoyados de la definición de Dolf Zillmann (1979), concluimos que las agresiones tienen como objetivo cumplir las metas del atacante, que es el infligir daño a otra persona o grupo de personas como recompensa a su conducta.

Aun cuando es evidente que las agresiones o actos de violencia son actitudes que se consideran antiguos, el término bullying se empieza a utilizar a partir de la década de los 70 por el Doctor Dan Olweus (1983). La etimología de la palabra bullying deriva del vocablo inglés “bull”, cuyo significado en español es “toro”, que se caracteriza por ser un animal fuerte e intimidador que puede arremeter fácilmente con animales más débiles (Monité, s.f.). En el idioma español, la palabra bullying, comúnmente se traduce utilizando los términos de intimidar o acosar.

Olweus empleó este término al encontrarse ante la necesidad de estudiar a fondo los fenómenos de agresiones que comenzaron a surgir en las escuelas, donde los estudiantes atacaban a los que consideraban débiles, con el único objetivo de infringirles temor, dolor o dañarlos. Finalmente, en la definición de Olweus se menciona que un estudiante es agredido cuando está expuesto de manera continua a situaciones de acoso por parte de uno o más estudiantes (Olweus,1983).

> "*Un estudiante es acosado o victimizado cuando está expuesto de manera repetitiva a acciones negativas por parte de uno o más estudiantes, sin capacidad para defenderse.*"

Recientemente, los investigadores Liu y Graves, en su definición puntualizan que una situación de bullying presenta las siguientes características principales, que permiten categorizarla como tal: (1) tienen la intención de lastimar (2) realizada de manera repetitiva, y (3) dirigida a una persona considerada tímida o un blanco fácil (Liu & Graves, 2011). Una vez explicado el concepto, para efectos de esta investigación, establecemos la definición de bullying como el comportamiento negativo, repetitivo e intencional, de una o más personas,

dirigido contra otra persona que tiene dificultad para defenderse, con la intención de intimidarla o dañarla.

2.1 Bullying

A lo largo de la literatura han existido diversas definiciones de bullying, todas ellas coincidiendo en que se trata de un abuso de poder realizado de manera repetitiva e intencional por parte del acosador hacia una persona que se encuentre de alguna forma en desventaja, con el fin último de causar algún daño (Continente et al., 2010). Aunque el bullying es un fenómeno que se ha presentado desde siempre, actualmente ha obtenido gran notoriedad debido al incremento de incidentes de violencia escolar y número de suicidios relacionados con este comportamiento. Es de destacar que esta práctica termina por dañar, de alguna manera, tanto a las víctimas y agresores, como a los observadores de la misma; corriendo, con mayor o menor medida, el riesgo de presentar síntomas de depresión, ansiedad, pensamientos suicidas, baja autoestima, entre otros (Loredo-Abdalá et al., 2008).

En consecuencia, a esta preocupante situación se han desarrollado programas de prevención de bullying, además de campañas contra el acoso escolar. El autor del término bullying, Olweus, fue uno de los pioneros en desarrollar programas de prevención para atender situaciones que empezaron a surgir en esa época. Por ejemplo, en el año de 1983, en Noruega, murieron por suicidio 3 estudiantes víctimas de la intimidación. El país tomó la decisión de iniciar una campaña nacional contra el acoso escolar, es así como se desarrolló el Programa de Prevención del Bullying de Olweus. Este programa se basa en la reestructuración del entorno social creado por los maestros, personal escolar, estudiantes y padres con la ayuda de psicólogos y consejeros (Olweus, 1997). Existen 4 objetivos que se desarrollan en este programa:

1. Aumentar la conciencia del acosador y la víctima sobre el problema.
2. Lograr una participación activa por parte de los maestros y los padres.
3. Desarrollar reglas claras contra la intimidación.
4. Proporcionar apoyo y protección a las víctimas.

Si bien el bullying se presenta con mayor frecuencia entre niños y adolescentes en etapa escolar, existen casos en los que se desarrolla dentro del área laboral, militar y recreativa (Loredo-Abdalá et al., 2008). Sin embargo, las acciones de atención a este fenómeno se han dado ya desde hace algunos años principalmente en el sector educativo. Por ejemplo, en nuestro país, con el fin de estudiar y llegar a comprender más a fondo la situación actual del fenómeno bullying, el *Instituto Nacional para la Evaluación de la Educación* (INEE), por medio de la *Dirección Nacional de Escuelas*, realizó un estudio sobre disciplina, violencia y consumo de sustancias nocivas a la salud, aplicado a estudiantes pertenecientes al nivel primaria y secundaria de la República Mexicana (García et al., 2007). Dicho estudio se llevó a cabo mediante dos exploraciones, la primera realizada en 2005 por medio de cuestionarios dirigidos a docentes y alumnos pertenecientes a los niveles educativos de interés, siendo complementado con los resultados obtenidos en los *Exámenes de la Calidad y el Logro Educativo* (Excale). Lo anterior, con la intención de comprender el funcionamiento de las instituciones escolares, así como alcanzar un entendimiento que permitiera explicar los

diferenciales del logro escolar. La segunda exploración del estudio se realizó en 2006 y comprendía entrevistas tanto de alumnos como de personal docente y directivo de escuelas secundarias. El objetivo de esta exploración fue establecer una correlación entre la organización, el funcionamiento y las problemáticas presentes en escuelas secundarias generales, técnicas y telesecundarias.

Del estudio realizado se concluyó que, de los 47,852 alumnos encuestados en nivel primaria, 34.8% participó en actos de violencia como perpetradores de agresiones físicas e intimidación hacia sus compañeros. De igual forma, se obtuvo que el 35.5% de los estudiantes de primaria han sido víctimas de alguna forma de intimidación y bullying. Por otro lado, de los 52,251 alumnos encuestados en nivel secundaria 38.6% ha participado en algún acto de violencia y bullying, mientras que el 39.4% aseguró ser víctima de alguna de estas prácticas (García et al., 2007).

En otro estudio más focalizado a situaciones de violencia, mediante la aplicación de la *1ra Encuesta Nacional Exclusión, Intolerancia y Violencia en Escuelas Públicas de Educación Media Superior* del año 2008, se muestran los porcentajes obtenidos del levantamiento de datos realizado en el año 2007 por el *Instituto Nacional de Salud Pública*. El conjunto evaluado comprendía 13,104 adolescentes de entre 15 y 19 años. El 44.3% de los hombres reportaron haber recibido insultos por parte de sus compañeros, mientras que el 23% de las mujeres reportó la misma actividad perpetrada por sus compañeros. Por otra parte, el 41.4% de los chicos y el 20.7% de las mujeres, reportó recibir apodos ofensivos por parte de sus compañeros (Szekely, 2008).

Más recientemente, en el año 2013, la *Subsecretaría de Educación Media Superior (*SEMS*),* de la *Secretaría de Educación Pública* (SEP), realizó la *Tercera Encuesta Nacional sobre Exclusión, Intolerancia y Violencia en las Escuelas de Educación Media Superior* con una muestra de 1500 estudiantes, donde se obtuvo que el 38.6% de los estudiantes han recibido algún tipo de insulto por parte de sus compañeros (SEMS & SEP, 2014).

Los índices de este fenómeno social son consistentes en el ámbito internacional. En el año 2014, durante un estudio que se llevó a cabo en Estados Unidos, los autores de (Juvonen & Graham, 2014) realizaron una revisión de los hallazgos relacionados con este complicado tema, esta investigación concluyó que entre el 20% y el 25% de los adolescentes están directamente relacionados con prácticas de bullying, ya sea actuando como los perpetradores de las agresiones, o siendo las víctimas que las padecen. De acuerdo con el *Departamento de Justicia de los Estados Unidos*, la *Oficina de Estadísticas de Justicia*, y la *Encuesta Nacional de Victimización del Crimen*, realizada en el año 2015, se identifica que el 20.8% aproximadamente, de 5,041,000 estudiantes de edades entre 12 y 18 años son víctimas de bullying (Lessne & Yanez, 2016). Al mismo tiempo, estudios realizados en (Oliveros et al., 2009), muestran que entre el 40% y 50% de los adolescentes de Perú y Colombia, afirman haber realizado prácticas de intimidación y bullying contra otros.

Debido a esta grave situación, los países, gobernantes y sociedad en general, realizan un examen de conciencia ante un fenómeno que no es nuevo y ha afectado a los estudiantes de los distintos niveles académicos desde siempre, con el objetivo de crear programas educativos anti-bullying. Por ejemplo, existen programas anti-bullying orientados a la concientización

social como el conocido por las siglas KIVA (Kiusaamista Vastaan en finés, que significa contra el acoso escolar), que es una metodología que resultó de una investigación en la Universidad de Turku, Finlandia. Este programa se basa en el cambio de actitud y respuesta de los espectadores antes una situación de bullying. Se sustenta en que el acoso escolar es un tema público común y concierne a todos los alumnos, no es solo un asunto privado entre el acosador y la víctima. Mediante este concepto se consigue que los niños adquieran conciencia de la gravedad y la responsabilidad de denunciar cualquier indicio que hayan presenciado de escenarios de bullying (Williford et al., 2012).

KIVA se continúa llevando a cabo en la actualidad en el 90% de las escuelas finlandesas y se ha adoptado en otros países como Holanda, Reino Unido, Francia, Italia, Suecia, España, entre otros (Martínez, 2020). El programa está basado en acciones universales que son perfectamente adaptables a cualquier región al incluir sesiones en clase, juegos online, reuniones con docentes, así como el proporcionar una retroalimentación a los participantes. Una de las acciones específicas que incluye el programa es que al detectar un caso de bullying se trate el tema directamente con la víctima y los perpetradores, además de invitar a los compañeros de clase a apoyar a las víctimas, así como mantener informados a los padres (Herkama & Salmivalli, 2018).

En otra iniciativa internacional, del instituto *Front Marítim de Barcelona*, se creó el programa TEI "Tutoría entre iguales", como técnica preventiva y de intervención para mejorar la convivencia escolar, con el propósito de involucrar a los niños y generar una dinámica de apoyo entre ellos en donde ninguna actividad de bullying tenga lugar. En este programa los alumnos participan como sujetos activos, sirviendo como tutores de alumnos más pequeños, incrementando la autoestima de los alumnos involucrados. Esta actividad permite reducir los niveles de inseguridad, facilitar el proceso de integración de los alumnos y compensar el desequilibrio de poder y fuerzas, mostrando TOLERANCIA CERO a actos de violencia o maltrato (Bellido, 2015).

Mas recientemente, en (Martínez, 2020) mencionan como el Ministerio de Educación del Gobierno de España realiza en su página Web una recopilación de los proyectos más importantes a nivel internacional para controlar y contrarrestar las actividades de bullying presente en el aula escolar. Entre ellos destacan programas como el ya mencionado KIVA, así como otros igual de efectivos, entre ellos se encuentran:

- *Mybullying*: Es un programa online que genera un mapa social del salón de clases y propone ciertas técnicas que tienen como objetivo evitar el aislamiento de estudiantes en desventaja o vulnerables.
- *NOHATE (No Hate Speech online)*: Es una campaña desarrollada por el Consejo de Europa que convoca a jóvenes para actuar en defensa de los Derechos Humanos y en contra de la intolerancia generada en Internet.
- *ARBAX (Against racial bullying and xenophobia)*: Programa financiado por la Unión Europea contra el racismo y la xenofobia.

En México se cuenta con *El Programa Nacional de Convivencia Escolar* (PNCE). Este programa enfatiza el fortalecimiento personal de los alumnos mediante técnicas psicológicas que promueven el crecimiento personal, como el apoyo en el fortalecimiento de la autoestima,

desarrollo de su autonomía, manejar sus emociones, resolución de conflictos aplicando el diálogo de manera asertiva, así como otras situaciones o desafíos que se les presenten en la vida cotidiana. Mediante técnicas individuales el PNCE promueve acciones que se emplean en toda la comunidad educativa con el objetivo de favorecer el desarrollo de docentes y fortalecer el liderazgo de directores, así como fomentar el mejoramiento del vínculo familiar y la relación escuela-familia; lo anterior con el objetivo de propiciar ambientes escolares pacíficos (Secretaría de Educación Pública, 2019).

Algunos puntos importantes que abarca la iniciativa del PNCE es el promover mediante la intervención pedagógica el que los y las alumnas reconozcan su valor propio; aprender a respetar a los demás y así mismos; aprendan educación emocional; manejen técnicas de comunicación; establezcan acuerdos; respeten las reglas y aprendan a manejar de manera asertiva cualquier tipo de conflicto que se presente en la vida cotidiana. Así como, favorece el desarrollo de docentes y directores con técnicas de convivencia escolar, impulsando la participación de las familias y coadyuvando en la prevención de situaciones de bullying (Secretaría de Educación Pública, 2019).

Adicionalmente a los programas desarrollados por distintos países, con el objetivo de prevenir y contrarrestar el bullying, también existen organizaciones encargadas de educar a la población en esta área y apoyar el desarrollo de nuevos programas como (Martínez, 2020):

- **Fundación ANAR** (Ayuda a Niños y Adolescentes en Riesgo), organización sin fines de lucro que se dedica a la defensa de los derechos de los niños y adolescentes en situación de riesgo y desamparo mediante proyectos desarrollados en España y Latinoamérica. Se originó en la Convención de los Derechos del Niño de Naciones Unidas en el año de 1970, sin embargo, aun en la actualidad sigue tomando acción en contra del acoso escolar.

- **Asociación Española para la Prevención del Acoso Escolar (AEPAE),** organización no gubernamental sin fines de lucro. Esta organización está comprometida con la prevención de bullying, estando constituida por psicólogos, educadores sociales, abogados, expertos en seguridad y familiares de víctimas de acoso escolar. Ha sido seleccionada tanto por la Comisión Europea como por el DIF (Dirección de Infancia y Familia) de México para revisar los programas de prevención de bullying. Esta asociación es la fundadora del Plan Nacional contra el Acoso Escolar de España que tiene como fin mejorar la convivencia en los centros educativos y el combatir situaciones de bullying. Se puso en marcha durante el ciclo escolar 2016/2017 en 32 colegios de España, disminuyendo la incidencia de víctimas de bullying en un 49%. Intervinieron 4,506 niños, niñas y adolescentes, reduciendo el número de víctimas de 403 a 208 casos.

- **Save The Children Fund** es una organización internacional no gubernamental fundada en 1919 que tiene como objetivo proteger los derechos de los niños, asegurar su supervivencia, educación y protección contra la violencia.

Aun cuando se han implementado programas anti-bullying en diversos países y niveles educativos, el bullying es un fenómeno evolutivo que ha adquirido nuevas formas de presentarse. En la actualidad se emplea una forma de bullying denominada cyberbullying que se desarrolla en entornos digitales mediante el envío de agresiones en forma de datos multimodales.

Esta problemática ha adquirido nuevas dimensiones, originadas por la inclusión de tecnologías de la información y la comunicación (TIC), como las redes sociales, celulares y el uso de Internet en general; ampliando significativamente el alcance y las categorías de las prácticas de bullying. En el ámbito educativo, el acoso ya no se limita a las horas de escuela mediante la agresión presencial, habiendo adquirido una modalidad que le permite al acosador realizar bullying a través de mensajes anónimos, llamadas, grabaciones, etc. Creando un escenario donde la víctima no tiene donde esconderse, y se encuentra expuesta ante una mayor cantidad de público (Benavides, 2015). Esta nueva forma de acoso recibe el nombre de cyberbullying.

El 42.8% de alumnos encuestados a nivel primaria y el 52.5% de nivel secundaria, consideran que la violencia se desarrolla dentro de la escuela, mientras que el 60.9% en primaria y el 60.7% en secundaria, derivado del uso de esta nueva tecnología, creen que este problema persiste fuera de la escuela (García et al., 2007). Al realizar un estudio con una muestra de 335,519 estudiantes en edades entre 12 y 18 años, en (Modecki et al., 2014) se encontró que el 35% de ellos había participado en un acto de bullying como perpetrador o víctima, y 15% se había relacionado en situaciones de cyberbullying. Se estima que aproximadamente 14.6 millones de adolescentes en Estados Unidos se han relacionado con prácticas de bullying y 6.2 millones han actuado como víctimas o perpetradores de cyberbullying (Giumetti & Kowalski, 2016). El estudio realizado en (Giumetti & Kowalski, 2016) permitió concluir que el 88% de los adolescentes involucrados en actividades de bullying presencial, que actúan como perpetradores o víctimas, se encuentran de igual forma extendiendo ese comportamiento a escenarios de cyberbullying.

2.2 Cyberbullying

El cyberbullying tiene los mismos efectos que el bullying tradicional, daña la confianza y la autoestima de la víctima, provocando ansiedad, frustración e ideas suicidas. Ante esta variación, los programas de atención de esta problemática y los desarrollos tecnológicos tienen que adaptarse para detectar a tiempo situaciones de cyberbullying mediante el uso de tecnologías de información y comunicación actuales, como las redes sociales, celulares e Internet; con el objetivo de actuar en contra de este tipo de agresión y proveer atención oportuna a la víctima.

En la búsqueda de contrarrestar esta nueva forma de acoso escolar se han creado diferentes propuestas tecnológicas. Por ejemplo, los estudiantes de la *Universidad Internacional de Valencia*, de Valencia España, desarrollaron 6 aplicaciones con el objetivo de frenar el bullying y acoso escolar. Las aplicaciones mencionadas se nombraron *MyFriends*, *Bully Freezone*, *Antibullying* Sage, *Antibullying-sector*, *Ray Chat* y *Treelp*, las cuales

proporcionan información y consejo sobre bullying mediante la reproducción de videos, pantallas interactivas, juegos y chats (Universidad de Valencia, 2015).

Mediante el uso de las TIC, específicamente haciendo uso de técnicas de Inteligencia Artificial (IA), se han realizado investigaciones a nivel internacional para detectar y atender los problemas que genera el fenómeno de cyberbullying. Por ejemplo, los autores de (Xu et al., 2012) en el año 2012, demostraron que el análisis de las publicaciones realizadas en redes sociales, mediante técnicas de procesamiento del lenguaje natural, permite reconocer y establecer una categorización del lenguaje utilizado en prácticas catalogadas como agresivas, generando la base requerida para identificar conversaciones en las cuales se estén desarrollando situaciones de bullying digital. A esta forma de cyberbullying, que se presenta mediante el envío de mensajes agresivos en redes sociales, se le conoce como acoso en línea.

Durante el año 2012, los investigadores de (Chen et al., 2012) se centraron en la detección de mensajes abusivos publicados en Twitter, que pudieran corresponder a situaciones donde los niños están siendo perpetradores o son víctimas de bullying. Los autores de este estudio desarrollaron un programa de monitoreo llamado *SocialFilter*, utilizando el método *Four-I*, consistente en identificar a los usuarios utilizando la red social, que permite concluir si los mensajes enviados por el usuario contienen lenguaje relacionado con actitudes de bullying. Mediante un modelo basado en *n-gramas*, que consiste en identificar patrones considerando múltiples combinaciones de palabras, se pudo detectar el efecto propagador de estos mensajes y proporcionar una retroalimentación a padres y maestros para que apliquen alguna solución al problema.

En el año 2013 se realizó un estudio donde se analizan las palabras frecuentemente utilizadas durante ataques de cyberbullying (Kontostathis et al., 2013); lo anterior mediante la recolección de mensajes utilizados en la red social *Formspring.me,* la cual opera por medio de la realización de preguntas entre usuarios. Los resultados de este análisis muestran que entre el 7% y el 14% de los mensajes analizados presentan contenido de cyberbullying. Este estudio logró determinar los términos comúnmente empleados en agresiones de cyberbullying, así como el contexto en el que son utilizados. Mediante aprendizaje máquina y la adquisición de nuevo conocimiento, lograron clasificar cada una de las preguntas realizadas entre usuarios a través de la identificación de la densidad de cyberbullying presente en el contenido.

Los investigadores de (Bretschneider et al., 2014) propusieron en el año 2014, un modelo de clasificación basado en la identificación de patrones para detectar mensajes con contenido abusivo. En su investigación determinaron relevante considerar el ruido presente en los textos escritos, representaciones codificadas de una palabra o coloquialismos, propios del estilo de comunicación utilizado por los usuarios al enviar mensajes, así como la identificación de frases relacionadas con los usuarios para el pre procesamiento de la información. Finalmente, el clasificador utiliza la información preprocesada para detectar patrones que conectan a una persona con manejo de lenguaje agresivo o palabras insultantes.

Por otra parte, en este mismo año, en (Margono et al., 2014), utilizaron minería de datos para identificar términos comunes de la lengua de indonesia. Palabras utilizadas frecuentemente para establecer actitudes de bullying, lo anterior mediante el análisis de

publicaciones realizadas en la red social Twitter. Dicho análisis les permitió determinar el léxico que se utiliza actualmente para generar ofensas y bullying.

En (Potha & Maragoudakis, 2015) utilizaron un conjunto de conversaciones reales entre perpetradores y víctimas, denotando de manera numérica la gravedad de los términos utilizados por el perpetrador y buscaron formas de representarlos mediante una descomposición de valor singular. Finalmente, mediante una red neuronal pronostican el nivel de insulto de una oración o pregunta, en relación con dos o tres de las oraciones previamente realizadas por el perpetrador, proporcionando un indicador inmediato de la gravedad de acoso o agresión dentro de una conversación.

MCDefender es un sistema móvil de defensa contra el bullying cibernético en redes sociales, desarrollado por los autores de (Vishwamitra et al., 2017). Este sistema detecta cyberbullying en Twitter y provee una intervención dinámica en incidentes detectados, utilizando como medio las plataformas móviles. En este proyecto, se estudia un mecanismo de detección en dos direcciones, al detectar cyberbullying en mensajes antes de que estos se envíen y ocultar aquellos que se reciban en el dispositivo móvil.

Los autores de (Chatzakou et al., 2017a) propusieron una metodología para detectar intimidación y acciones de cyberbullying en publicaciones de Twitter. Dicha metodología extrae textos, datos de los usuarios y atributos de la red, y analiza las propiedades o características de los agresores, identificando aquellas que los diferencian de los usuarios regulares. Concluyeron en algunas de las características presentes en los perpetradores, como su presencia en la red social, su nivel de participación y su nivel de aceptación por parte del grupo social o popularidad.

Investigaciones como las presentadas en (Del Bosque & Garza, 2016; Katsure, 2015), han generado modelos de aprendizaje automático con el objetivo de predecir textos agresivos en redes sociales, principalmente en Twitter, mediante la implementación de diferentes técnicas como redes neuronales, máquinas de soporte vectorial, sistemas difusos y regresión lineal.

Actualmente, con la intención de detectar cyberbullies y comportamiento agresivo en usuarios de Twitter, en (Chatzakou et al., 2017b) entrenaron clasificadores de aprendizaje máquina con características obtenidas de usuarios etiquetados como *bullies*, dichas características incluían, el contenido y lenguaje utilizado para comunicarse vía redes sociales. El clasificador *Random Forest* obtuvo el mejor resultado con una precisión de 92.2%, demostrando que la metodología y técnicas utilizadas permiten una alta detección de usuarios *bullies*.

Investigaciones descriptivas sobre la presencia de cyberbullying, similares a las descritas anteriormente, se encuentran en (Chatzakou et al., 2017c) donde analizan un conjunto de tweets relacionados con Gamergate, una campaña de cyberbulliyng generada en Twitter, que comenzó como denigración a las mujeres en la industria del juego, eventualmente convirtiendose en amenazas de violencia, violación y asesinato; y (Raisi & Huang, 2016) donde proponen un modelo para descubrir acosadores y víctimas de bullying. En esta investigación identificaron que el método propuesto puede detectar nuevo vocabulario de

intimidación a partir de un corpus de interacciones y un diccionario de indicadores de bullying.

Mientras que en (Bradley & Kendall, 2019) desarrollaron *At-Risk for Middle School Educators*, una simulación por computadora creada por *Kognito Interactive,* que tiene como objetivo capacitar de manera efectiva a docentes de nivel secundaria sobre prácticas efectivas de identificación y derivación de estudiantes que estén posiblemente en riesgo de bullying. Estudiantes pertenecientes al programa de Maestría en Artes en la Enseñanza (MAT), en el estado de Nueva York, complementaron *At-Risk for Middle School Educators* mediante encuestas y participación de una discusión en línea sobre la capacitación con dicha simulación. Los resultados de esta investigación arrojaron que mejoró significativamente la probabilidad de que los maestros recomienden servicios psicológicos a un estudiante que se encuentra en una situación compleja. Otro resultado destacable es la mejora significativa en la confianza y la competencia de los maestros para abordar y canalizar a los estudiantes en riesgo a profesionales de salud mental.

La investigación realizada en (Gada et al., 2021) presenta varias aplicaciones basadas en un modelo de clasificación que utiliza word2vec, y una arquitectura LSTM-CNN (por sus siglas en inglés de long short-term memory y convolutional neruonal network) para clasificar tweets y comentarios con cyberbullying. En una primera aproximación aplican el modelo desarrollado en una página web obteniendo resultados óptimos, por lo que adicionalmente se planeaba implementar dicho modelo en Telegram Bot, para emitir advertencias a los usuarios que van a generar un ataque de cyberbullying; si se producen cierto número de advertencias, el usuario sería eliminado. Posteriormente, crearon una extensión de Chrome que censura el lenguaje obsceno en cualquier página predeterminada, eliminando o resaltando palabras altisonantes. Finalmente, crearon una extensión de WhatsApp Chrome para dar seguimiento a las conversaciones de WhatsApp y desactivar el botón de envío, además de mostrar una advertencia al usuario que está generando un comentario desagradable.

En (Perera & Pumudu, 2021) describen un sistema que tiene como objetivo detectar cyberbullying de forma automática mediante la detección de texto que contenga características de acoso cibernético, además de otros temas como racismo, contenido de acoso sexual y palabras ofensivas. Los autores de esta investigación consideran que el texto utilizado en ataques de cyberbullying ha cambiado con el tiempo y que en la actualidad se utilizan palabras con una connotación altamente ofensiva o como ellos lo nombran palabras extremas. En esta investigación utilizaron técnicas tradicionales de extracción de características, como la frecuencia de términos y frecuencia de ocurrencia del término en la colección de documentos (TF-IDF), N-gramas y el análisis de sentimiento.

2.3 Investigaciones relacionadas y modelado

La detección automática de *cyberbullying* entre adolescentes es un tema atendido ampliamente por investigadores de las áreas de procesamiento de lenguaje natural y aprendizaje automático (Rosa et al., 2019). La explotación de redes sociales, como *Twitter y YouTube* (Dinakar et al., 2012), ha detonado la generación de grandes bancos de datos para

este fin. Estas fuentes de información son comúnmente utilizadas para generar modelos de aprendizaje automático y aprendizaje profundo que permiten detectar situaciones de acoso entre grupos de jóvenes. Sin embargo, la selección de la fuente y los mecanismos de preprocesado utilizados para la creación de estos bancos de datos no siempre son los adecuados. Diversos estudios han utilizado datos de redes sociales que permiten una fácil obtención, por ejemplo el presentado en (Xu et al., 2012), que incluyen interacciones entre personas de diversos círculos sociales y edades. En otras investigaciones se construyen corpus usando diccionarios de palabras vulgares y no coloquiales que permiten identificar comentarios que contienen palabras similares (Núñez-Prado et al., 2020; Aragón et al., 2019; Aragón et al., 2020). Mediante estas técnicas filtran *tweets* con interacciones de interés (presencia de bullying). Desafortunadamente, el contexto de interacción de estas fuentes no es el utilizado por grupos cerrados de amigos o compañeros de estudio, que son los escenarios donde se presentan situaciones de *cyberbullying* con mayor frecuencia e intensidad entre estudiantes (Cerezo, 2009).

Considerando este patrón por parte de los investigadores en utilizar fuentes de datos poco representativas de los sujetos de estudio de interés, en (Rosa et al., 2019) se presenta un análisis detallado de 22 estudios, realizados de 2011 a 2018, orientados a la detección automática de *cyberbullying*. En esta revisión de literatura se asevera que la atención a este fenómeno generalmente está mal representada, derivando en modelos que difícilmente pueden ser trasladados al mundo real. Entre las principales causas identificadas se encuentra la ausencia de métodos uniformes de evaluación de los modelos propuestos y la inconsistencia en el origen, preprocesado y etiquetado de los datos utilizados para entrenar dichos modelos. Destacando que en la mayoría de los estudios no se especifican los detalles de creación del corpus usado para el modelado. En la Tabla 1 se presentan las principales características evaluadas en el estudio (datos extraídos de (Rosa et al., 2019)).

Tabla 1. Características de 22 corpus utilizados en el uso de modelos de Aprendizaje Automático y Aprendizaje Profundo para tratar cyberbullying: 2011-2018

Característica	Sí	No	NP
Diálogos entre compañeros	1	20	1
Etiquetadores expertos en el tema	1	7	14
Obtención usando barrido web	21	0	1
Idioma inglés	20	2	0
Agresiones de un solo mensaje	19	3	0

* NP- corresponde a información no proporcionada

Adicionalmente, en el estudio se identifican como las principales fuentes de información las redes sociales *Twitter* y *YouTube*, obteniendo los datos mediante un proceso rápido de barrido Web o utilizando un API público. También, se encontró que únicamente el 13% de los estudios consideran tres de los siguientes cuatro criterios de etiquetado: (1) lenguaje hostil, (2) intención de afectar una tercera persona, (3) repetición del comportamiento, y (4) ataque entre compañeros. Un 13% reportó que considera sólo uno de

los criterios y el otro 74% ninguno de ellos. Adicionalmente, se enfatiza la ausencia de diálogos grupales en los corpus utilizados, con un 86% de diálogos analizados con mensajes agresivos aislados, lo cual evita la identificación de reincidencia. Finalmente, en ningún caso se hace referencia a la manipulación de datos en varios formatos (multimodales) y la mayoría de los corpus son en el idioma inglés.

Durante 2019 y 2020 se identificaron 13 nuevas investigaciones atendiendo este fenómeno social, las cuales nuevamente enfatizan la red social *Twitter* como fuente principal de datos, y conversaciones en el idioma inglés. Adicionalmente, solo una de estas investigaciones considera datos multimodales como parte del corpus, una de ellas texto e imágenes y una más texto y emoticonos. Pero en ninguna de ellas se presentan detalles mayores sobre el preprocesado y etiquetado de los textos (Balakrishnan et al., 2019; Banerjee et al., 2019; Tapia et al., 2018; Chelmis, & Yao, 2019; Kumar, & Garg, 2019; Mouheb et al., 2019; Cheng et al., 2019; Chen, Guo et al., 2019; Cheng, Li et al., 2019; Yao et al., 2019; Samghabadi et al., 2020; Wang et al. 2020; Fortunatus et al., 2020; Van Bruwaene et al., 2020) (ver Tabla 2).

Tabla 2. Características de 13 corpus utilizados en la modelos de Aprendizaje Automático y Aprendizaje Profundo para tratar cyberbullying: 2019-2020

Característica	Sí	No	NP
Diálogos entre compañeros	1	12	0
Etiquetadores expertos en el tema	3	5	5
Obtención usando barrido web	12	1	0
Idioma inglés	10	2	1
Agresiones de un solo mensaje	10	2	1

* NP- corresponde a información no proporcionada

Adicionalmente, se reportan diferencias significativas en el desempeño de los modelos de clasificación generados, y discrepancia en la métrica utilizada para su evaluación (ej. exactitud, precisión, sensibilidad y Valor-F). Por ejemplo, mediante el uso del clasificador de Bosques Aleatorios (RF, por sus siglas en inglés), eficiente con grandes bancos de datos, (Rosa et al., 2019) reporta una medida de Valor-F de 0.59 con un corpus de 2,999 *tweets* y de 0.45 con un corpus de 13,260 textos, usando este mismo clasificador (Balakrishnan et al., 2019) utilizó un corpus de 9,484 *tweets* y reporta un Valor-F de 0.92.

En (Samghabadi et al., 2020) se evaluó el desempeño del algoritmo de Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés), reportando un Valor-F de 0.76 con un corpus de 297 conversaciones. En (Chelmis & Yao, 2019) utilizaron Regresión Logística (LR, por sus siglas en inglés) con un corpus de 10,000 comentarios, reportando un Valor-F del 0.86.

Los autores de (Kumar & Garg, 2019), usando un corpus de 8,000 *tweets* reportan una exactitud de 0.74 para RF y de 0.64 mediante un clasificador Bayesiano Ingenuo (NB, por sus siglas en inglés). Con respecto a algoritmos de aprendizaje profundo, utilizando una Red Neuronal Convolucional (CNN, por sus siglas en inglés), redes usadas principalmente en

modelos para la clasificación de imágenes, y recientemente en el ámbito de NLP, (Banerjee et al., 2019) reporta una exactitud de 0.93 con un corpus de 69,874 tweets. En la Tabla 3 se resume lo encontrado en estos estudios.

Tabla 3. Nivel de desempeño de algoritmos de clasificación

Fuente	Algoritmo	Corpus	Valor-F	Exactitud
[7]	RF	2,999 tweets	0.59	
[7]	RF	13,260 textos	0.45	
[14]	RF	9,484 tweets	0.92	
[23]	SVM	297 conversaciones	0.76	
[17]	LR	10,000 comentarios	0.86	
[18]	RF	8,000 tweets		0.74
[18]	NB	8,000 tweets		0.64
[15]	CNN	69,874 tweets		0.93

Los resultados que refleja esta revisión de literatura muestran una baja atención en el uso de bancos de datos representativos, tanto al seleccionar la fuente de datos del corpus como al determinar las características del contenido y su etiquetado. Así como, diferencias significativas en el nivel de desempeño obtenido, ya sea por variaciones en el tamaño del corpus utilizado, como por los algoritmos de clasificación implementados; destacando un mejor desempeño al utilizar CNN.

En otra revisión de literatura, presentada recientemente en (Elsafoury et al., 2021), se ratifica que persiste la baja atención en las características mencionadas al crear corpus para el análisis de cyberbullying, y se menciona un problema adicional sobre el desbalance entre las muestras positivas (con *bullying*) y negativas (sin *bullying*). Reportando que, de un total de 24 corpus analizados, la mayoría presenta un nivel alto de desbalance, con niveles de hasta 10% de muestras positivas. Solo una investigación reporta un balance aceptable con 42% de muestras positivas. Esta desproporción en las muestras del corpus implica sesgos al momento de entrenar a los modelos, ya que se complica distinguir presencia o ausencia de *cyberbullying* (ver Tabla 4).

Tabla 4. Balance de muestras positivas (con-bullying).

% del total de los Corpus analizados	Positivas (Con Bullying)	Negativas (Sin-Bullying)
30%	10%	90%
36%	11%-29%	89%-71%
30%	30%-39%	70%-61%
4%	42%	58%

* Porcentajes con base a 24 corpus analizados en (Elsafoury et al., 2021).

Con base en lo anterior, se enfatiza que, en el estudio reportado en la presente investigación, se le da una atención especial a la creación de un corpus con contenido de cyberbullying representativo, el cual permite generar modelos con datos que garantizan la factibilidad de trasladarse a entornos reales de interacción social. Se presenta como producto

un corpus con las siguientes características: fuente de datos representativa de jóvenes estudiantes de nivel medio superior y superior, diálogos con conversaciones entre 2 o más participantes, balance en textos positivos y negativos, entre otras.

Marco teórico

El fundamento teórico que sustenta nuestra investigación se basa principalmente en las siguientes áreas de estudio que atiende el proceso de aprendizaje automático: 1) *Creación de un Corpus*, 2) *Preprocesamiento de los datos, y 3*) *Modelos de Clasificación*. A continuación, se describen los elementos a considerar en nuestra investigación derivados de cada de una de estas áreas de conocimiento.

2.4 Características de los datos usados en Cyberbullying

El cyberbullying se ha convertido en la actualidad en un tema social que ha adquirido relevancia debido al aumento de casos. Resaltando la importancia de la detección oportuna y la aplicación de estrategias que permitan contrarrestar sus efectos. Aunado a esto se ha adquirido gran popularidad entre las investigaciones científicas por el amplio espectro de oportunidades en el análisis de esta problemática, y la disponibilidad de datos abiertos a los que se pueden acceder con relativa facilidad. Sin embargo, mientras los avances computacionales incrementan y los recursos son cada vez más accesibles, existen restricciones que limitan el progreso en la detección de cyberbullying, como lo es el acceso a conjuntos de datos de calidad que sean representativos de la población de estudio.

Por ejemplo, en (Emmery et al., 2021) evalúan las dificultades de la recopilación de datos de calidad o representativos, así como un análisis de los recursos públicos disponibles. Destacan que los datos a los que se tiene acceso comúnmente pertenecen a redes sociales donde no se requiere algún acceso interno, sin embargo, resaltan que la práctica de cyberbullying en entornos reales se lleva a cabo principalmente “a puertas cerradas” en conversaciones privadas entre grupos de personas.

Por cuestiones de seguridad y confidencialidad, no es posible en la actualidad tener acceso a este tipo de conversaciones privadas, es por eso que en la práctica las investigaciones tienen que trabajar con conjuntos de datos de libre acceso que en su mayoría exhiben un alto grado de sesgo entre los mensajes positivos y negativos, ausencia de diálogos o conversacioes largas y baja representatividad en contenido, por lo que no capturan con precisión el lenguaje utilizado y las características lingüísticas generalizables con presencia de cyberbullying en una conversación proveniente de una audiencia objetivo.

En atención al problema de calidad y representatividad de los corpus, con el objetivo de obtener resultados más precisos en la clasificación de cyberbullying en (Larochelle & Khoury, 2020), experimentaron creando un conjunto de datos uniendo otros que cumplieran con alguna definición del concepto general de cyberbullying. Para llevar a cabo esta

investigación analizaron ocho conjuntos de datos. Los conjuntos de datos seleccionados son recopilados de distintas plataformas, con diferentes formatos (ver Tabla 5).

Tabla 5. Número de mensajes, palabras únicas, y conteo.

Conjunto de datos	**Mensajes**	**Palabras Únicas**	**Total de Palabras**
A-	3,002	10,989	34,606
A+	14,841	23,193	155,359
B-	7,957	15,153	73,973
B+	3,627	9,531	40,027
C-	8,619	14,445	104,946
C+	556	2,554	8,049
D-	2,898	13,195	56,848
D+	1,049	4,377	13,183
E-	7,002	12,468	61,023
E+	877	3,706	10,273
F-	114,677	162,524	4,410,757
F+	12,979	32,001	398,342
G-	1,358,749	284,164	38,232,678
G+	85,150	63,441	2,032,318
H-	49,155	99,083	1,983,200
H+	6,463	21,748	235,820

*Tabla tomada de (Larochelle & Khoury, 2020).

En una primera fase dividieron el conjunto de datos en textos positivos representando a aquellos con presencia de cyberbullying (A+) y negativos a los textos con ausencia de cyberbullying (A-). Aun cuando en (Larochelle & Khoury, 2020) no dan mayor detalle estadístico que los datos mostrados en la Tabla V, podemos observar el comportamiento heterogéneo de frecuencias entre palabras, e identificar que la cantidad de palabras únicas o poco frecuentes presentes en los textos es proporcional a mayor número de textos y total de palabras (ver Tabla 6); independientemente del origen de los datos. Estos aspectos son un indicativo de valores atípicos que afecta el desempeño de los modelos de clasificación debido a la presencia de palabras con errores ortográficos u otros errores gramaticales.

Tabla 6. Número de textos, conjunto de palabras.

Conjunto de datos	Número Textos	Número Palabras	Palabras Únicas	%Palabras Únicas
A	17,843	189,965	34,182	18%
B	11,584	114,000	24,684	22%
C	9,175	112,995	16,999	15%
D	3,947	70,031	**17,572**	**25%**
E	7,879	71,296	16,174	23%
F	127,656	4,809,099	194,525	4%
G	1,443,899	40,264,996	**347,605**	**0.86%**
H	55,618	2,219,020	120,831	5%

Adicionalmente, podemos observar que cada uno de los conjuntos de datos analizados en la investigación de (Larochelle & Khoury, 2020) muestra un alto grado de desbalance. En la Tabla 7 podemos observar que la mayoría de los conjuntos muestra una mayor cantidad de mensajes sin presencia de bullying y una cantidad no representativa de mensajes con bullying. Solo uno de los conjuntos de datos muestra un desbalance a favor de las clases positivas de cyberbullying. Lo anterior descrito describe corpus que sesgan el entrenamiento de cualquier modelo de clasificación. Una cantidad tan limitada de mensajes con presencia de bullying no refleja la realidad de un conjunto de datos representativo de una población objetivo, adicionalmente un desbalance tan significativo complica la capacidad de los modelos de clasificación para distinguir entre una clase y otra.

Tabla 7. Balance de muestras positivas (con-bullying).

Conjunto de datos	Positivas (Con Bullying)	Negativas (Sin-Bullying)
A	**83%**	**17%**
B	31%	69%
C	6%	94%
D	27%	73%
E	11%	89%
F	10%	90%
G	6%	94%
H	12%	88%

2.5 Creación del Corpus

Existen distintas definiciones para el termino *Corpus*, en lingüística un corpus es un conjunto de textos que representan el uso del lenguaje real para una determinada situación, almacenados en un medio informático (Quiroz, 2007). El autor de (Parodi, 2008) propone 3

condiciones para definir un conjunto de datos como un corpus: 1) que su contenido consista de textos provenientes de situaciones reales; 2) su recolección debe de ser guiada y seguir ciertos parámetros para su posterior análisis, y debe ser documentada de forma que permita la posibilidad de replicar el proceso de acopio, así como, mantener una extensión considerable de acuerdo al contexto del corpus, claridad y confiabilidad; y 3) debe de estar disponible en algún formato electrónico que permita su análisis mediante modelos y algoritmos computaciones.

Con base en lo anterior, uno de los objetivos primordiales al crear un corpus es el definir los criterios que determinen la representatividad de la lengua y la temática en cuestión (Pérez-Hernandez, 2002). Siendo un tema en discusión el determinar la relevancia del corpus con respecto a su calidad (representatividad) y cantidad (tamaño). En cuanto al muestreo requerido para el análisis de redes sociales, resulta complicado determinar el número de muestras necesarias para representar las relaciones existentes entre actores de una población; que, aun cuando se determine un número de actores o participantes significativo, no es una garantía de la representatividad de sus interacciones (Pericas & Olive, 1999).

En el caso particular de la presente investigación, la generación del corpus para la identificación de cyberbullying en entornos educativos de habla hispana del nivel de estudio objetivo, se define mediante la recolección y categorización de un conjunto de diálogos en lenguaje español con presencia de mensajes con diferentes niveles de “agresividad” o “acoso”, provenientes de conversaciones en redes sociales (tales como Facebook, WhatsApp, Snapchat, Instagram, etc.), entre grupos de jóvenes de los niveles educativos de preparatoria y universidad. El corpus generado tiene como objetivo permitir el análisis lingüístico (derivado de los mensajes textuales), paralingüístico (derivado de emoticones que representan expresiones de sentimiento, tales como felicidad, tristeza, llanto, etc.), y visual (por imágenes y memes utilizados), que permitan la generación de modelos que puedan apoyar a identificar situaciones de acoso o agresión.

2.6 Preprocesamiento de los datos

El Internet se ha convertido en una fuente inagotable de información con una gran cantidad de datos al alcance de un clic. Cada vez es más fácil y rápido obtener grandes bases de datos y corpus mediante barridos web, tanto de redes sociales (Twitter principalmente) como de otros sitios. Sin embargo, la calidad de los datos en su mayoría se encuentra comprometida al presentar factores desfavorables como ruido, inconsistencias, valores perdidos, entre otros.

En particular, los datos de texto obtenidos de redes sociales tienen una gran cantidad de ruido, derivado del tipo de interacción que se han apropiado los usuarios de estos sitios. Lo anterior se origina de la singular manera de escribir de los usuarios, donde se pueden presentar combinaciones de letras y números para representar una palabra, errores ortográficos, emoticonos, entre otras maneras del uso informal del lenguaje.

El objetivo de realizar tareas de preprocesamiento de datos es preparar los datos crudos o con ruido para el modelado, mediante análisis y transformaciones, para obtener un cuerpo de

datos que será más útil de manejar al realizar tareas analíticas, minería de datos o procesamiento del lenguaje natural (Pyle, 1999). El preprocesamiento de datos consiste en una serie de pasos, los cuales son aplicados de acuerdo a las características de cada investigación en particular y dependen del objetivo a analizar si son aplicables o no, pero generalmente se incluyen categorías generales de tokenización (identificación de elementos básicos del lenguaje), normalización y sustitución.

En (Mayo, 2017) proponen un framework de preprocesamiento de texto que engloba las tareas generales que deben de aplicarse a un conjunto de datos de este tipo. El framework mencionado consiste de un proceso de tareas recurrentes como se muestra en la Figura 1.

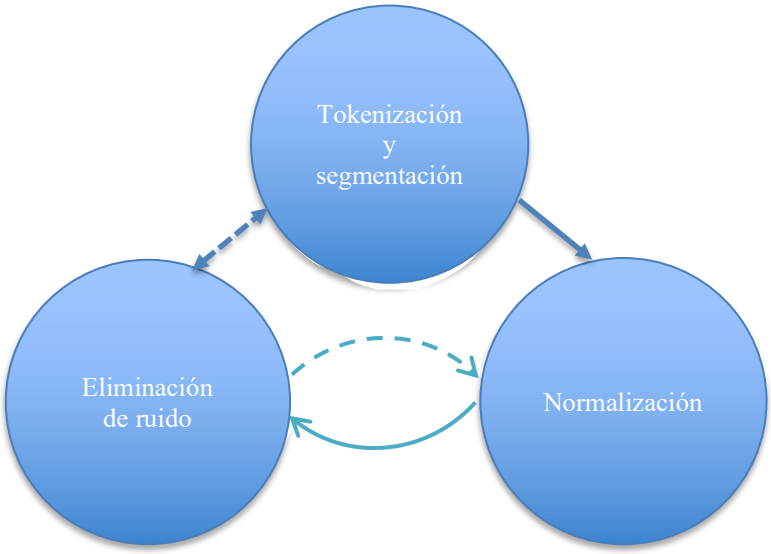

Figura 1. Framework para procesamiento de datos de texto (Mayo, 2017).

En la fase de tokenización se divide una cadena de texto en partes más pequeñas denominadas tokens. A esta fase también se le conoce como segmentación de texto o análisis léxico. La segmentación hace referencia a la descomposición de una gran cantidad de texto en partes más pequeñas sin llegar a ser palabras, es decir, la descomposición de un texto en párrafos u oraciones, mientras que el término tokenización se reserva para el proceso de descomposición exclusivamente resultante en palabras (tokens).

La fase de normalización consiste en todas aquellas tareas que permitan regularizar en un formato o estilo de igualdad de condiciones, por ejemplo; convertir el texto en minúsculas o mayúsculas, eliminar signos de puntuación, convertir números en sus equivalentes de palabras, eliminar espacios en blanco o eliminar stop words (palabras de poca utilidad conocidas como palabras vacías). Las stopwords son aquellas palabras que carecen de un significado por si solas como, pronombres, preposiciones y artículos.

Por último, la fase de eliminación de ruido consiste en todas aquellas tareas que sean requeridas para la limpieza del texto y que va a depender directamente de la fuente de datos de donde se han obtenido. Es decir, si el texto se ha obtenido de una página web, se necesitarán eliminar ciertos metadatos provenientes del formato de los archivos, tales como HTML, XML,

etc.; por otra parte, si los datos se han obtenido de archivos de texto tal vez sea necesario eliminar, encabezados o pies de página (Mayo, 2017).

La revisión de literatura realizada por los autores de (Salawu, 2020) muestra que las tareas de preprocesamiento mayormente realizadas por 22 de las investigaciones actuales son: tokenización y steaming (obtención de la forma base de las palabras), conversión de mayúsculas a minúsculas, eliminación de stop words, eliminación de caracteres especiales o sustitución, y corrección de ortografía y gramática, entre otras.

El autor (Saluwu, 2020) menciona que a pesar de que la etapa de preprocesamiento es un paso estándar de limpieza de datos y que es casi idéntico en la mayoría de los trabajos de investigación presentes en la literatura actual, es una tarea que debe depender del conjunto de datos seleccionado y de la propia tarea a realizarse, así como de los modelos utilizados. Por ejemplo, las técnicas actuales de corrección de ortografía y gramática son aplicables a conjuntos de texto en el idioma inglés, ya que la mayoría de las herramientas disponibles solo se encuentran disponibles para este idioma.

2.7 Modelos de Clasificación

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial que consiste en automatizar la identificación de patrones mediante algoritmos. El objetivo principal es desarrollar técnicas que permitan que los sistemas o máquinas aprendan así como la creación de modelos que permita resolver un problema o tarea dada, por ejemplo, generando recomendaciones de preferencias en alguna aplicación, mediante la clasificación de textos positivos o negativos, entre otros tipos de aplicaciones (Simeone, 2018).

Un algoritmo utilizado en aprendizaje automático es aquel proceso computacional que, mediante una entrada de datos, previamente preparados o etiquetados, logra realizar un modelo que realiza una tarea deseada sin haber sido programado para ello. En una etapa inicial, al algoritmo se le proporcionan muestras de datos y se configura para producir un resultado deseado, para que posteriormente pueda generalizar y producir nuevos resultados, ya sea recomendaciones o categorizaciones, a partir de datos de entrada no conocidos. A este conjunto de etapas se le denomina aprendizaje (El Naqa & Murphy, 2015).

El aprendizaje automático emula la forma en que los humanos y otras criaturas aprenden. En la vida diaria el humano realiza un sin número de procesos de aprendizaje automático, con actividades tan simples como distinguir entre manzanas y peras.

Existen tres categorías diferentes para clasificar el aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. En el primero de ellos, aprendizaje supervisado, los algoritmos trabajan con datos previamente etiquetados en categorías predefinidas. Por otro lado, en el aprendizaje no supervisado los datos utilizados para generar el modelo no cuentan con una categoría o etiqueta. Finalmente, el aprendizaje por refuerzo no dispone de datos etiquetados o no etiquetados, el algoritmo de aprendizaje consiste en registrar el resultado de realizar una acción al interactuar con el mundo exterior. Es decir, el

algoritmo tiene la capacidad de distinguir un resultado favorable o desfavorable, aprendiendo mediante un esquema basado en el ensayo y error. El conjunto de datos utilizado para el modelado en la presente investigación se encuentra en la categoría de aprendizaje supervisado, ya que se cuenta con datos previamente etiquetados con presencia o ausencia de bullying (categorización binaria); entre otras categorías complementarias. Con base en lo anterior, a continuación, se describen los algoritmos más utilizados en esta categoría de aprendizaje automático.

2.7.1 Aprendizaje Supervisado

Como se comentó anteriormente, en el aprendizaje supervisado los algoritmos son entrenados con datos etiquetados, así es como los modelos generados aprenden a asignar una etiqueta de salida a un nuevo valor que recibe como entrada (Simeone, 2018). Este tipo de aprendizaje se utiliza para resolver problemas de clasificación o predicción, realizando tareas como identificación de dígitos, detección de fraudes, predicciones meteorológicas, clasificación de correos electrónicos no deseados (spam), entre otros. Algunos de los algoritmos más populares y con mejor desempeño utilizados en el aprendizaje automático son descritos a continuación:

2.7.1.1 Regresión Logística

La Regresión Logística Binaria es un modelo estadístico que se utiliza para conocer la relación existente entre una variable dependiente dicotómica, es decir, una variable que solo puede tomar un valor entre dos posibles, y una o más variables independientes llamadas covariables que pueden llegar a ser cuantitativas o cualitativas. El objetivo de este modelo es predecir una clase binaria y obtener una estimación ajustada de la probabilidad de que ocurra un evento a partir de múltiples variables predictoras (Berlanga Silvente & Vilá Baños, 2014).

A este modelo se le atribuyen 3 objetivos principales:

- El cuantificar la importancia de la relación entre la variable dependiente y las covariables.
- Busca clarificar la interacción entre las covariables y la variable dependiente.
- Y finalmente, clasificar nuevos valores dentro de las categorías de la variable dependiente.

2.7.1.2 Máquinas de vectores de soporte (Support Vector Machine)

Support Vector Machines o Máquinas de Vectores de Soporte, es un algoritmo de aprendizaje automático que se utiliza para clasificación y regresión de espacios no lineales que tiene la capacidad de separar un conjunto de observaciones en dos clases mediante un hiperplano de separación (Farhadian et al., 2020).

Si bien en un inicio se pensó como un método de clasificación binaria, se ha extendido su aplicación a la resolución de problemas de clasificación múltiple y de regresión. Se considera como uno de los mejores clasificadores en aprendizaje automático y aprendizaje estadístico, ya que pueden ser utilizados para la resolución de una gran cantidad de problemas.

El objetivo de este algoritmo es encontrar un hiperplano en un espacio de dimensiones N (el número de características) que separe de la mejor forma posible dos clases diferentes de datos. Se busca el hiperplano que tenga el margen máximo y más amplio entre las dos clases, es decir, la distancia máxima entre los puntos de datos de ambas clases (Farhadian et al., 2020).

El margen está definido por la máxima anchura de la región que no tiene puntos de datos interiores paralelos al hiperplano, maximizarlo permite un incremento de la confianza en la clasificación de los datos (ver Figura 2).

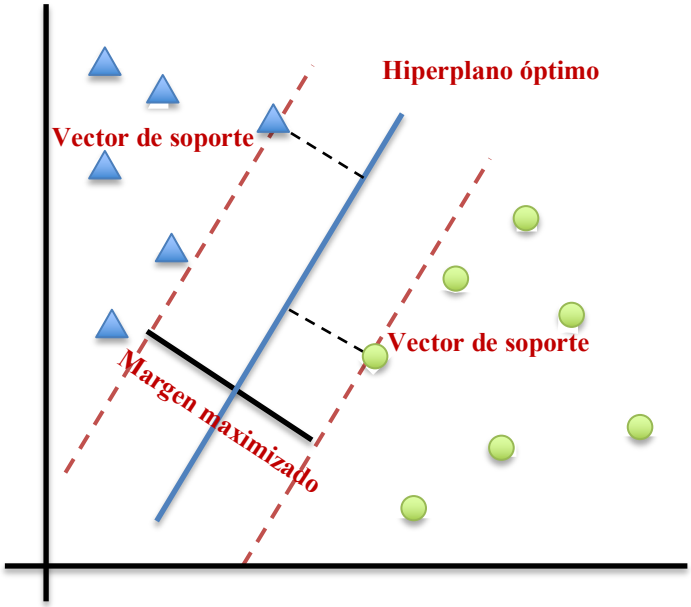

Figura 2. Máquina de vector de soporte (SVM).

El algoritmo sólo puede encontrar este hiperplano en problemas que permiten separación lineal. Sin embargo, al pertenecer a un algoritmo de ML de tipo núcleo o kernel, estima una medida de similitud entre dos vectores en el espacio mapeado. Lo anterior simplifica los límites de decisión no lineales para hacerlos lineales en el espacio dimensional de características, con la expectativa de que resulte más fácil separar las clases después de esta transformación. Esto se conoce como truco de kernel (Mathworks, s.f.).

Dentro de las ventajas de utilizar este tipo de clasificadores es que logran un buen rendimiento en tareas de clasificación y regresión; es muy eficaz con muestras de entrenamiento de grandes dimensiones, incluso cuando el número de dimensiones es mayor al número de muestras. Adicionalmente, aun cuando están formulados para clasificación binaria, se puede implementar un SVM multiclase al combinar varios clasificadores binarios; es

eficiente en memoria y el uso de kernels los hacen más flexibles y capaces de gestionar problemas no lineales (Mathworks, s.f.).

2.7.1.3 Bósques Aleatorios (Random Forest)

Los Bosques Aleatorios o Random Forest usan un modelo de aprendizaje automático supervisado que permite resolver problemas tanto de clasificación como de predicción. Este tipo de modelado busca incrementar el potencial del algoritmo tradicional de Árboles de Decisión, aumentando su desempeño y controlando el sobreajuste durante el entrenamiento. En concreto, los bosques aleatorios están constituidos por muchos árboles de decisión (bosque) y proporcionan un resultado en función de las predicciones individuales de los árboles de decisión, realizando un promedio de sus predicciones. Por lo tanto, cada árbol depende de un vector aleatorio que se prueba de manera independiente durante el entrenamiento y da como resultado una clasificación o predicción de los árboles individuales (Ho, 1995; Ho, 1998).

El método Random Forest tiene como base una modificación del proceso llamado *bagging* que de-correlaciona los árboles que se han generado durante el proceso del algoritmo. El bagging promedia un conjunto de modelos reduciendo su varianza solo en el caso que se cumpla la no correlación. Es decir, Random Forest asegura que el comportamiento de cada árbol individual no esté relacionado con el comportamiento de otro árbol en el modelo (ver Figura 3).

Durante el Bagging el bosque aleatorio permite que cada árbol individual tome muestras aleatorias del conjunto de datos con reemplazo (*Bootstrap Aggregation*), lo que permite que cada árbol sea diferente. Esta técnica no subdivide los datos de entrenamiento, si el conjunto de datos es de tamaño N se toma una muestra de tamaño N con reemplazo (Ho, 1998).

Un ejemplo simple: si tenemos el conjunto de datos [A, B, C, D, E, F], podríamos tomar una muestra [A, B, B, C, F, F] donde observamos que se permite repetir datos de entrenamiento, seleccionados al azar, generando un conjunto de datos que le damos a alguno de nuestros árboles (muestreo con reemplazo).

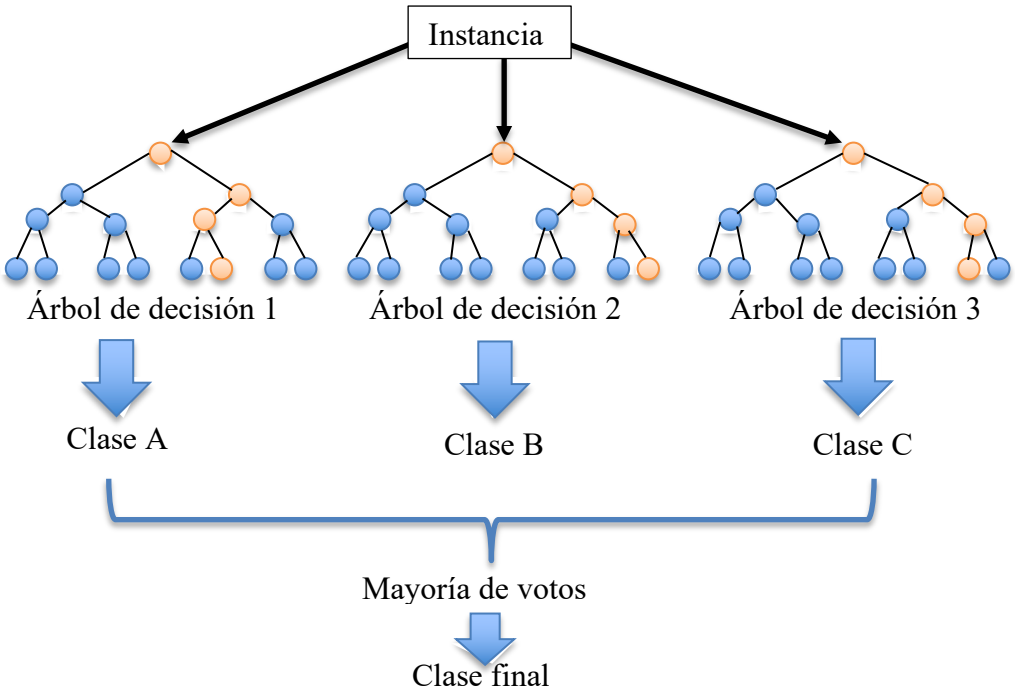

Figura 3. Representación del proceso de clasificación de Bosque aleatorio con tres árboles de decisión como predictores

Estos tipos de clasificadores son fáciles de interpretar, útiles en la exploración de los datos, no requieren un preprocesamiento exhaustivo de los datos y pueden manejar tanto predictores cuantitativos como cualitativos.

2.7.1.4 Clasificador Naive Bayes

Naive Bayes, es un modelo de clasificación supervisado y probabilístico basado en el Teorema de Bayes con el supuesto de que existe independencia de las variables predictoras, de ahí recibe el apelativo de ingenuo. Es decir, la presencia o respectiva ausencia de alguna de las características de los datos es independiente y no está relacionada con la ausencia o presencia de cualquier otra (Pacheco Leal et al., 2005).

Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

Ecuación del Teorema de Bayes en la que se basa el modelo se muestra a continuación (Larragaña et al., 1997):

$$(1) \qquad P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Donde:

P(A) y P(B) son las probabilidades a priori de los sucesos A y B.
Siendo A y B dos sucesos aleatorios cuyas probabilidades se denotan por p(A) y p(B) respectivamente, verificando que $p(B) > 0$.

P(B|A) es la probabilidad condicionada del suceso B dado el suceso A.
P(A|B) es la probabilidad a posteriori del suceso A conocido que se verifica el suceso B.

El clasificador Naive Bayes es muy fácil de implementar, proporciona una ventaja fácil y rápida de clasificar problemas binarios o multiclase. Se considera mejor en la clasificación de datos independientes (Pacheco Leal et al., 2005).

2.7.1.5 Aprendizaje profundo y redes neuronales

Una subcategoría del aprendizaje automático y la Inteligencia Artificial es el Aprendizaje Profundo o *Deep Learning*, técnica de aprendizaje que imita la forma en que los humanos obtienen cierto tipo de conocimiento (Gegundez Arias & Pérez Borrego, 2021). Una característica destacable de este subcampo es la utilización de diferentes estructuras de redes neuronales, similar a la interacción entre neuronas de un humano, que permiten un aprendizaje mucho más significativo sobre los datos. El Término profundo o *deep* hace referencia a la cantidad de capas de representación que usa el modelo.

Una red neuronal artificial es la representación de un modelo matemático el cual pretende simular el comportamiento biológico de las neuronas y la estructura que forman en el cerebro. Consiste en un conjunto de unidades, ***neuronas*** que se conectan entre sí para permitir la transmisión de señales. Las señales viajan a través de la red neuronal, donde se llevan a cabo los procesamientos computacionales y las operaciones requeridas, dando como resultado ciertos valores de salida.

Una red neuronal puede consistir de una o más capas de neuronas. En la primera capa de cualquier red neuronal se toma como entrada el conjunto de datos del que partimos sin ningún procesamiento, seguido de esto se procesan y se extrae la información correspondiente, esta información se pasa a la siguiente capa como parámetros de entrada. Mediante un proceso iterativo se obtiene una predicción que se puede mejorar cuando la red neuronal ajusta los hiper parámetros (Torres, 2018).

Los algoritmos de *deep learning* siguen por lo general los siguientes pasos:

1. Recopilar un conjunto de datos asociados a un problema.

2. Diseñar una función de coste apropiada al problema, también conocida como función de pérdida (*loss function*)
3. Seleccionar un modelo de red neuronal y establecer sus hiper parámetros (tamaño, características, etc.)
4. Aplicar un algoritmo de optimización para minimizar la función de coste ajustando los parámetros de la red.

En el aprendizaje profundo se consideran varios tipos de redes neuronales, tales como las redes neuronales recurrentes, conocidas por su capacidad para procesar datos secuenciales, y las redes neuronales convolucionales utilizadas ampliamente en el procesamiento de imágenes. A continuación, se describen algunas de sus características.

Redes Neuronales Recurrentes

Las Redes Neuronales Recurrentes son sistemas dinámicos utilizados mayormente para el reconocimiento de voz, análisis de video y el procesamiento del lenguaje natural. Son capaces de realizar una gran variedad de tareas computacionales que incluyen el tratamiento de secuencias, modelación de sistemas dinámicos y la continuidad de trayectorias de predicción no lineal (Sak et al., 2015).

Son conocidas como redes *espacio-temporales* o dinámicas, no se utilizan tan solo para clasificar un dato en particular, sino que también tienen la capacidad de generar nuevas secuencias. Permiten estimar lo que ocurrirá en el futuro, gracias a su arquitectura y a cómo trabajan, son capaces de ser creativas, tienen la habilidad de predecir. Un ejemplo simple de cómo actúan es al predecir cuál es la siguiente nota en una melodía y así poder componer una pieza musical (Sak, Senior et al., 2014).

Las Redes Neuronales Recurrentes se diferencian de otras al no limitarse en recibir una entrada (*input*) y a partir de este obtener una salida (*output*), que se pasará a la siguiente neurona, sino que envían el output no solo a la neurona siguiente, sino también a sí misma, de modo que se convierte en un input para la misma neurona, de aquí su nombre de recurrencia. Cada neurona recurrente recibe dos tipos de inputs, el de la neurona anterior y el propio que viene del estado anterior, de esta forma la neurona se auto retroalimenta, lo que permite que haya una cierta memoria y tenga la capacidad de recordar estados pasados hasta una cierta cantidad de momentos en el tiempo, llamados épocas (Sak, Vinyals et al., 2014)

Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales son un tipo de red neuronal que tienen como objetivo simular el funcionamiento de un cerebro biológico al trabajar de manera muy similar a las neuronas en la corteza visual primaria, por lo que son muy efectivas para tareas de visión artificial, como en la clasificación y segmentación de imágenes, entre otras aplicaciones (Pérez-Carrasco et al., 2011).

Este tipo de redes neuronales consisten de múltiples capas de filtros convolucionales de una o más dimensiones, después de cada capa se añade una función de mapeo causal no lineal.

Existen una serie de pasos empleados por este tipo de red neuronal para realizar trabajos de clasificación y son las siguientes fases:

1. Extracción de características compuestas de neuronas convolucionales.
2. Reducción por muestreo.
3. Generación de neuronas perceptrón sencillas, que realizan la clasificación sobre las características extraídas.

Este tipo de redes aprenden a reconocer y diferenciar objetos dentro de imágenes, mediante el proceso de aprendizaje de un algoritmo de entrenamiento que se basa en las características únicas de los objetos y la generalización (Araujo et al., 2018).

Capítulo 3

Metodología

El enfoque metodológico en que se sustenta esta investigación se basa en la metodología para la implementación de modelos de aprendizaje automático propuesta en (Elsafoury et al., 2021), el flujo de trabajo del aprendizaje automático (***Machine Learning Pipeline)***, que consiste en una serie de pasos ordenados que establecen la secuencia de trabajo de aprendizaje automático. Los pasos mencionados en esta metodología son los siguientes (Elsafoury et al., 2021):

1) Recopilación de datos: En este paso del pipeline se define el origen y anotación de los datos; los datos son procesados en un formato establecido que permita su utilización en pasos posteriores, el decir el formato con el que se presentarán los datos para ser utilizados con los modelos de clasificación en aprendizaje automático. Complementariamente se realiza una validación de los datos, identificando estadísticas, rangos, datos atípicos, así como anomalías.

2) Preprocesamiento de los datos: Este paso es de vital importancia, implica preparar los datos mediante técnicas de limpieza para obtener un conjunto de datos final que sea adecuado para la utilización de modelos de aprendizaje automático, ya que el utilizar datos sin preprocesar puede generar resultados incorrectos.

3) Selección de características: En esta etapa se realiza un etiquetado y reducción de los datos. Además de la selección de las características más importantes del conjunto de datos que nos permitirán un mejor rendimiento en el entrenamiento de modelos.

4) Entrenamiento de modelos: El entrenamiento de los modelos es la tarea principal del pipeline, en esta etapa el modelo se entrena para tomar el conjunto de datos preprocesados y dar como resultado una salida con la mayor precisión posible.

5) Evaluación de modelos: Finalmente en esta etapa final del pipeline, se determina el rendimiento del modelo, se realizan cálculos para determinar métricas como precisión, exactitud, etc. que permite identificar el conjunto óptimo de parámetros que dará como resultado el modelo final.

Para esta investigación, la recopilación de datos (origen y anotación de datos), parte fundamental de este estudio, consiste en la definición, recopilación y tratamiento de los datos representativos del dominio de interés, en este caso datos con presencia de Cyberbullying. Por otro lado, los pasos de preprocesamiento de datos, selección de características y entrenamiento de modelos son primordiales en el modelado de conocimiento, definiendo las estructuras y algoritmos que permiten crear el modelo de clasificación que propone esta investigación. Y,

por último, el paso de evaluación de modelos del pipeline hace referencia a la medición del desempeño de los modelos generados y en su caso el desarrollo tecnológico del ambiente que permita evaluar el impacto del modelo seleccionado con usuarios reales.

A continuación, se describe los detalles de cada una de las faces seguidas en la metodología.

3.1 Recopilación de datos de Cyberbullying

Simultáneamente, para la generación del corpus con contenido de situaciones de cyberbullying, se realizó una adaptación de la metodología utilizada ampliamente en el área de estudio de minería de datos *Descubrimiento de Conocimiento en Bases de Datos* (KDD, por sus siglas en inglés) (Maimon & Rokach, 2015). La metodología implementada estuvo constituida por las cuatro etapas principales de KDD, que se describen a continuación:

1) *Selección de la población de interés*

En esta etapa de la metodología KDD, se pretende desarrollar un entendimiento sobre el dominio de estudio, todo descubrimiento previo sobre el dominio y la definición del objetivo es primordial para enfocar el significado del modelado y de la investigación.

2) *Acopio de datos*

El objetivo de esta etapa es seleccionar el conjunto de datos relevantes para nuestra investigación, ese conjunto de datos originales que nos permitirá resolver el problema objetivo. En esta etapa es importante homogeneizar los datos obtenidos y eliminar variables irrelevantes para que en procesos posteriores sea más fácil procesar y analizar.

3) *Preprocesamiento*

En el preprocesamiento de los datos se realiza una limpieza de los datos, se realiza la toma de decisiones con respecto a los criterios de identificación y tratamiento de los valores atípicos presentes en los datos, esos valores que causan ruido y afectan negativamente en el desempeño descriptivo o predictivo de los modelos generados. También se realiza un proceso de normalización de las variables presentes en el conjunto de datos.

Esta etapa generalmente no es atendida con la importancia que se requiere, sin embargo, es una de las etapas con mayor relevancia dado que los valores fuera de rango (valores atípicos), combinaciones fuera de contexto en el corpus, entre otros datos erróneos, generados intencionalmente o inconscientemente, pueden llevar a entorpecer el proceso de aprendizaje automático de los algoritmos y así conseguir resultados que no reflejan el comportamiento real.

4) *Transformación.*

La etapa de transformación nos permite procesar el conjunto de datos obtenido, con el fin de detectar las características útiles y los patrones evidentes, así como las relaciones entre ellos, para representar los datos de la mejor manera acorde a las características del modelado objetivo.

3.2 Definición operacional de las variables

Considerando las hipótesis planteadas, particularmente la H1 que supone sobre la existencia de factores que gestan la materialización de cyberbullying que son identificables y medibles, se buscó obtener información sobre factores adicionales (variables independientes, alternas al solo análisis de palabras) al contenido de las conversaciones que ayudaran a identificar con mayor precisión la variable dependiente (presencia de cyberbullying).

Variables independientes:

En el estudio se buscó analizar el impacto que factores como los siguientes (en sus respectivas categorías) tienen en la aparición de cyberbullying de forma individual o combinada, o alguno adicional identificado durante la investigación.

Exposición a medios electrónicos:

- Tiempo de uso
- Horario de uso
- Tipo de aplicaciones utilizadas
- Rol principal en aplicaciones (activo, pasivo)
- Cantidad de cyber-contactos
- Tipo de información que comparte (fotos, videos, audios, texto)
- Etiquetado de información (likes)
- Publicaciones en redes sociales

Comportamiento de grupos de interacción en las conversaciones:

- Cantidad de participantes
- Cantidad de hombres y mujeres
- Tópico, categoría y forma de agredir

Factores del entorno (interacción y cantidad de contactos), con los roles de:

- Padres
- Familiares
- Compañeros de escuela
- Amigos

Mediante el uso de los factores descritos anteriormente, se planteó analizar el desempeño que tienen los distintos modelos generados mediante el uso de variantes en las técnicas de clasificación usadas en IA tales como:

- Redes bayesianas
- Árboles de decisión
- Redes neuronales

Tipo de estudio y diseño general

La presente investigación se sustentó en un método de estudio cuantitativo, ya que se analizaron e implementaron técnicas para el modelado de clasificación del fenómeno social de cyberbullying, evaluando y comparando cuantitativamente su desempeño al utilizar datos con diversos formatos; cuantificando su nivel de confianza. Por lo anterior, analizaremos su nivel de:

- ***Confiabilidad***, nivel de confianza en identificación de cyberbullying.
- ***Eficiencia de rendimiento***, tiempo de respuesta y utilización de recursos.
- ***Escalabilidad***, al manipular diversos formatos de datos.

Paradigma de la Investigación

Se define como investigación experimental ya que se evaluó la precisión máxima alcanzable por parte de diferentes algoritmos de modelado y la combinación de variables independientes para la clasificación de comentarios con contenido agresivo. Se supuso que el comportamiento es influenciado por la fuente de datos proveniente de espacios de interacción representativos del campo social de interés.

Métrica para medición de desempeño de algoritmos

Una vez realizada la evaluación de las características de los datos a utilizar (dataset), la implementación de los modelos, y proceder a evaluar los resultados obtenidos, se debe determinar qué tan efectivo ha sido el modelo utilizado bajo alguna métrica de evaluación.

Existen distintas métricas que permiten evaluar la eficacia y rendimiento de los modelos de clasificación, es por lo tanto una elección importante la selección de la misma para comparar los modelos de aprendizaje automático.

Para realizar la evaluación de desempeño de los modelos generados se utilizaron las medidas de exactitud, precisión, sensibilidad y Valor-F. Sin embargo, se determinó utilizar la métrica de Valor-F (F-score) como medida principal al comparar los modelos generados, ya que esta prueba determina un valor único ponderando las métricas de precisión y sensibilidad y es utilizada generalmente en la prueba de algoritmos de recuperación de información, así como de clasificación de documentos. Las fórmulas de estimación se describen a continuación:

Precisión

La métrica de Precisión **(en inglés "Precision")** nos permite medir la **calidad** de un modelo de aprendizaje automáticoutilizado en tareas de clasificación. Se refiere a la dispersión del conjunto de datos a partir de mediciones repetidas de una magnitud. Cuando la disperción es menor la precisión es mayor (Rivera et al., 2018).

Se puede definir de una manera práctica al denotarla como el ***porcentaje de casos positivos detectados.*** Dado que estamos evaluando nuestros datos por su desempeño de predicciones "positivas".

Se representa por la proporción de verdaderos positivos dividido entre todos los resultados clasificados como positivos (tanto verdaderos positivos, como falsos positivos).

$$(2) \qquad precision = \frac{TP}{TP + FP}$$

Donde TP (True Positives) representa los verdaderos positivos, TN (True Negatives) para los verdaderos negativos, FP (False Positives) representando los falsos positivos y finalmente FN (False Negatives) para los falsos negativos.

La precisión es un gran estadístico, pero solo es útil cuando se tienen «datasets» simétricos.

Exactitud

La métrica de Exactitud **(en inglés, "Accuracy")** está definida como el porcentaje de casos que el modelo ha acertado. Esta se refiere al porcentaje de instancia correctamente clasificadas, Positivas o negativas, frente al total. En otras palabras, la Exactitud es la ***cantidad de clasificaciones positivas que fueron correctas*** (Rivera et al., 2018)**.**

Está métrica está relacionada con el sesgo de una estimación, donde si el sesgo es menor la estimación es más exacta.

Sin embargo, no se aconseja su uso en tareas de clasificación cuando se cuenta con un conjunto de datos donde las clases están desbalanceadas, ya que no es una métrica que permita medir de una manera válida el rendimiento de un modelo. Por otro lado, es recomendable en conjuntos de datos donde las clases están casi equilibradas.

$$(3) \qquad accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensibilidad

A la métrica sensibilidad también se le conoce como Recall y está definida como el cálculo denotado por el número de predicciones positivas correctas dividido por el número total de positivos.

La sensibilidad es la proporción de los casos positivos que fueron correctamente identificados por el modelo. También se le conoce como **Tasa de Verdaderos Positivos (True Positive Rate) ó TP,** si su resultado es uno entonces se han encontrado todos los verdaderos positivos en el dataset, por lo que no existe ruido; si su valor es cero los datos no poseen relevancia (Rivera et al, 2018).

$$(4) \qquad recall = \frac{TP}{TP + FN}$$

Esta métrica nos permitirá identificar si no se están perdiendo positivos.

Valor-F

La métrica de Valor F o puntaje F1 está definida como el promedio ponderado de precisión y sensibilidad, tomando en cuenta tanto falsos positivos como falsos negativos.

En la práctica no se espera calcular la precisión y la sensibilidad cada vez que creamos un modelo cuando nos enfrentamos a un problema de clasificación; por lo que el Valor F nos permite obtener una puntuación única que represente ambas variables (Rivera et al., 2018).

$$(5) \qquad F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Esta métrica es muy utilizada cuando se tiene una distribución de clases desbalanceada o desigual, y es muy empleada porque nos resume la precisión y sensibilidad en una sola métrica.

Capítulo 4

Diseño y desarrollo del corpus de datos con presencia de Cyberbullying y modelos de clasificación

4.1 Metodología para la creación del corpus

En la presente investigación se consideró la relevancia de contar con un corpus de datos proveniente de redes sociales representativas de los sujetos de estudio objetivo; definiendo como tal a adolescentes estudiantes mexicanos de nivel medio superior y superior. Incluyendo conversaciones donde interactúan un distinto número de participantes, y que a su vez incluyera datos multimodales. Como resultado, se creó un corpus con diálogos con presencia de cyberbullying proveniente de grupos usando las redes sociales cerradas de Facebook y WhatsApp.

El objetivo de la generación del corpus fue el de proveer de un banco de datos que permita realizar un análisis lingüístico procedente de conversaciones representativas de un entorno real de acoso escolar. Aportando la consideración de una característica paralingüística al contener emoticonos que representan expresiones de sentimiento, (tales como felicidad, tristeza, llanto, etc.), y aspectos visuales al asociarlo con imágenes y memes (imágenes con texto embebido). Lo anterior, con el propósito de generar modelos que permitan identificar situaciones de acoso o agresión basadas en datos más representativos al grupo social objetivo.

Para la generación del corpus propuesto, como se comentó en la sección anterior, se siguió una variante de la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD, por sus siglas en inglés) (Maimon & Rokach, 2005). La metodología implementada estuvo constituida por las cuatro etapas principales de KDD que se describen a continuación: (1) selección de la población de interés, (2) acopio de datos, (3) preprocesamiento, y (4) transformación.

4.1.1 Selección: Identificación de la población de interés

Aun cuando la práctica de acoso presencial y virtual se presenta en diversas áreas, como la laboral, recreativa o educativa, este fenómeno social se presenta con mayor frecuencia entre niños y adolescentes en etapa escolar (Loredo-Abdalá et al., 2008); particularmente en los niveles medio superior y superior (INEGI, 2014; Kocatürk & Türk-Kurtça, 2020; Reisen et al., 2019). Por lo tanto, en este trabajo se determinó utilizar como sujetos de estudio, grupos de jóvenes de los niveles educativos de preparatoria y universidad.

A diferencia de ejecutar un proceso KDD convencional, donde generalmente se cuenta con una base de datos previamente recopilada para proceder a seleccionar los datos relevantes y prioritarios a minar (Maimon & Rokach, 2005), en el presente trabajo se procedió a identificar una fuente de datos representativa de los sujetos de estudio y posteriormente a

recopilar los datos de interés, proceso descrito a continuación. Considerando que las fuentes usadas convencionalmente (principalmente Twitter) para obtener el corpus se asumía que no eran representativas de los entornos privados de interacción de los sujetos de estudio objetivo, se procedió a realizar una encuesta para determinar la preferencia de uso de las mismas. Tal como se describe a continuación.

En esta fase se realizó una encuesta a 158 estudiantes procedentes de 3 instituciones de educación media superior y superior. El grupo de participantes consistió en 82 hombres y 76 mujeres, con una edad promedio de 18 años. En el instrumento utilizado se preguntó por la red social de su preferencia para interactuar con sus amigos o compañeros de escuela, la cantidad de horas invertidas en esta actividad y el horario de uso. Los resultados de la encuesta se muestran en la Tabla 8, siendo evidente que la representatividad de Twitter, red social comúnmente utilizada en otros estudios, es mínima o casi nula en este nivel educativo. El documento de la encuesta aplicada se encuentra en el ***Apéndice A***.

Tabla 8. Redes sociales de preferencia por estudiantes de nivel medio superior y superior.

Red Social	Preferencia de uso
Snapchat	39%
Facebook	37%
Instagram	15%
WhatsApp	6%
Twitter	2%
Ninguna	1%

Con base en los resultados de esta encuesta, se determinó crear el corpus de datos con diálogos provenientes de grupos creados en tres de las redes sociales identificadas con mayor uso por jóvenes de los niveles educativos objetivo: Facebook, WhatsApp e Instagram. Snapchat no fue considerada debido a su característica funcional de auto borrado de conversaciones grupales después de 24 horas de haberse realizado.

4.1.2 Acopio de conversaciones

Se procedió a la creación del corpus objetivo invitando a participar voluntariamente a estudiantes de tres grupos escolares de aproximadamente 45 estudiantes cada uno. Bajo consentimiento expreso, un total de 40 estudiantes, pertenecientes a la población de interés, colaboró proporcionando un conjunto de diálogos en idioma español donde ellos consideraban se contaba con la presencia de diferentes niveles de “agresividad” o “acoso”. En esta fase del estudio se contó con la participación de 5 hombres del nivel medio superior y 16 del nivel superior, así como de 4 y 15 mujeres; respectivamente. Para el acopio de los datos, se estableció como requisito que la procedencia de las conversaciones fuera entre miembros de grupos privados registrados en las redes sociales previamente identificadas con mayor audiencia.

Con la finalidad de agilizar y hacer práctico el proceso de acopio, los datos se recibieron mediante archivos de texto, generados por las mismas aplicaciones de las redes sociales, o a través de capturas de pantalla; lo que fuera más fácil de compartir para los sujetos participantes en el estudio. En total se lograron obtener 472 diálogos provenientes de conversaciones grupales, realizadas en redes sociales donde interactuaron más de 200 jóvenes. Un total de 420 conversaciones fueron categorizadas con presencia de algún tipo de agresión, representando el 89% del conjunto de datos acopiado, seleccionado por el equipo de investigadores y posteriormente validados en el proceso de etiquetado por tres psicólogas expertas en lenguaje utilizado por los jóvenes. Tal como se esperaba, las conversaciones se obtuvieron de 2 de las redes sociales más representativas: Facebook y WhatsApp. Los detalles cuantitativos sobre el contenido del corpus de diálogos con cyberbullying se presentan en la Tabla 9.

Tabla 9. Concentrado de diálogos con presencia de cyberbullying.

Diálogos	Líneas de texto	Imágenes	Memes
420	3,114	55	31

4.1.3 Preprocesamiento: Pre-filtrado y transcripción de conversaciones

En esta etapa, similar a la fase de preprocesamiento de KDD, donde se analiza la calidad de los datos y se eliminan datos atípicos, se procedió a identificar conversaciones con cyberbullying y transcribirlas mediante un formato homogéneo para facilitar su posterior etiquetado y procesamiento. Para esto se tomaron en cuenta las siguientes consideraciones.

En la literatura se indica que una oración usada para ofender o agredir consiste de componentes claramente identificables, estos son: 1) dirección, es decir a quién va dirigido el insulto; y 2) la palabra o frase que, aun sin ser grotesca, puede llegar a ofender (Satapathy et al., 2015). Estas palabras o frases pueden ser clasificadas de acuerdo con el menor o mayor grado de ofensa que puede generarse al ser utilizadas. A continuación, se describen las principales categorías y el tipo de palabras o frases negativas o insultantes que resultan en un nivel de agresión (Satapathy et al., 2015):

- Peyorativas: aquellas palabras o frases que indican una idea desfavorable o despectiva.
- Obscenas: palabras o frases ofensivas al pudor, generalmente de contenido sexual.
- Profanas: Palabras o frases irrespetuosas, conocidas también como palabras fuertes o groserías.

Para el filtrado inicial de las conversaciones se utilizó esta categorización, considerando un diálogo con presencia de cyberbullying a aquel con comentarios con una o más palabras o frases negativas o insultantes y una dirección de ofensa.

Adicionalmente, en esta fase, se procedió a realizar la siguiente actividad de limpieza de conversaciones: eliminado de diálogos incompletos, descartado de conversaciones que no presentaran un escenario de una interacción real, o eliminado de diálogos con ausencia de agresiones. Al identificar las conversaciones con presencia de cyberbullying se procedió a la transcripción de los diálogos a archivos de texto para facilitar su posterior etiquetado, procesado y análisis. Siempre garantizando el anonimato de los participantes y la consistencia de los datos mediante el uso de identificadores únicos. Enfatizando que la transcripción de las conversaciones se realizó en forma literal, manteniendo consistencia en errores ortográficos y tipográficos, modismos, anglicismos, énfasis, entre otros.

4.1.4 Transformación: Etiquetado y categorización de conversaciones y enunciados

En la fase de transformación y reducción de datos en un proceso KDD se procede a definir mecanismos de identificación adecuado a cada tipo de datos (ej. uso de categorías en lugar de valores continuos) y reducción de dimensiones (ej. eliminando o reduciendo datos o registros incompletos) (Maimon & Rokach, 2005). Lo anterior permite obtener una representación de las características útiles para el proceso de minado.

En esta etapa del estudio, considerando nuestro objetivo de modelado, primeramente, se procedió a identificar datos en las conversaciones que no estuvieran en formato texto, tales como imágenes, memes o emoticonos. Para cada uno de ellos se definió un esquema de representación que fuera lo suficientemente descriptivo para mantener la comprensión y fluidez de lectura de una conversación. El diálogo presentado en la Figura 4 ejemplifica una conversación entre 3 usuarios, incluyendo referencias a imágenes (enunciado 1), otros usuarios no presentes en la conversación (enunciado 2) y emoticonos (enunciado 3):

@usuario1 [img419]
@usuario1 No wey y el otro día que fui a dar un rol con el [@usuario5] en baica
@usuario1 [emoticono mano-arriba]
@usuario2 Jajajaja
@usuario3 A no eras tu???? Jaja es que me confundi.. jaja es que estan identicos jajaja
@usuario1 A defendiendo a su bato…

Figura 4. Ejemplo de diálogo de un grupo de 3 estudiantes

Adicional a los archivos de texto obtenidos en el paso anterior, se generó un instrumento electrónico para el etiquetado de los diálogos que permitió categorizar las conversaciones utilizando tres niveles de profundidad o granularidad: general, descriptiva y por enunciado. El instrumento se incluye en el ***Apéndice B***:

a) *Nivel General*, cada diálogo fue categorizado por número de participantes y su sexo, tipo de interacción (uno a uno, uno a muchos, y muchos a muchos), y su

nivel de agresión (usando una escala Likert de 0 a 5, donde 0 representa sin agresión y 5 muy agresivo).

b) *Nivel Descriptivo*, el documento de etiquetado permitió la categorización de diálogos considerando el tópico de la conversación (Escolar, Recreativo, Familiar, Sociedad, Otro), la categoría de bullying (Racial, Sexual, Religioso, Apariencia, Desempeño académico, Nivel social, Otro), y el medio o la forma en que se realizaban los ataques (Rumores, Amenazas, Comentarios, Exclusión, Compartir información confidencial, Otro). Los tópicos, categorías y formas de agresión derivan de lo expuesto en la literatura (Kurniasih et al, 2020; Ansary, 2020; Berdugo-Gómez, 2020; Ruiz- Ramírez et al., 2020; Coob & Marín, 2020), con el aval de 3 profesionales en el área de psicología educativa.

c) *Nivel por enunciado*, finalmente, cada enunciado en las conversaciones fue clasificado con base en su nivel de agresión, utilizando la misma escala Likert referida en la categorización general. Adicionalmente, se solicitó identificar el enunciado detonador de la agresión.

El proceso de etiquetado se llevó a cabo por una terna de psicólogas especializadas en el área de cyberbullying y con experiencia en el lenguaje utilizado por los jóvenes de la población objetivo. En la Figura 5 se presenta un ejemplo del documento de etiquetado, donde se resalta el etiquetado de nivel de agresión por cada uno de los comentarios en una conversación, así como los apartados para indicar factores adicionales como cantidad de participantes, tópico, nivel de interacción, entre otros. Para evitar que el proceso de etiquetado se viera afectado por sesgos derivados de influencia entre opiniones, cada una de las etiquetadoras evaluó el total de los diálogos de forma independiente.

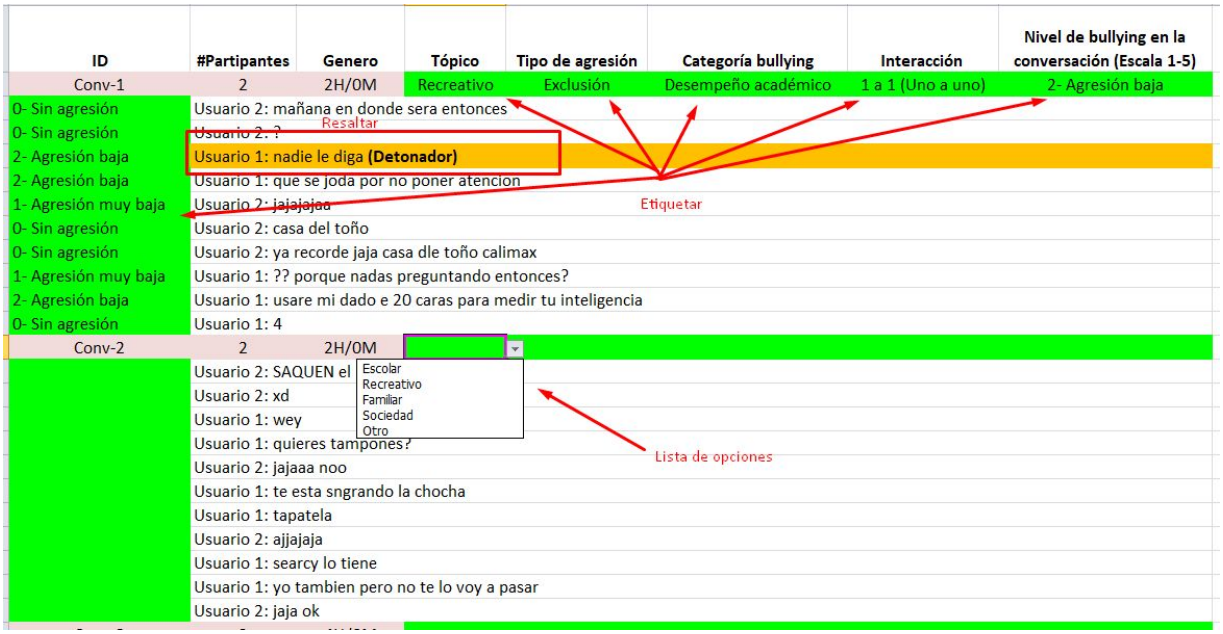

Figura 5. Ejemplo del instrumento de etiquetado a nivel general, descriptivo y por enunciado. El documento completo se encuentra el Apéndice B

Posterior al proceso de etiquetado, se generó un solo documento que concentró el etiquetado final de las conversaciones. La principal diferencia encontrada entre las categorizaciones realizadas por las etiquetadoras fue al determinar el nivel de agresión presente en diálogos o enunciados. Con el objetivo de determinar un nivel de acuerdo común entre las etiquetadoras se siguieron los siguientes criterios de homogeneización:

1) Concordancia entre las 3 etiquetadoras, se tomó la evaluación común.
2) Concordancia entre 2 etiquetadoras, se tomó la evaluación común mayor.
3) Discrepancia entre las 3 etiquetadoras, se realizó un promedio entre las evaluaciones.

Finalmente se obtuvo un corpus lingüístico basado en texto consistente de conversaciones con presencia de cyberbullying, etiquetado por expertas en el área. En la siguiente sección se describen las características generales de este corpus de datos.

4.1.5 Descripción de los documentos generados del corpus resultante

Como resultado de la metodología seguida y los datos obtenidos se construyó el conjunto final de archivos que conforman el corpus con presencia de situaciones reales de cyberbullying. El corpus resultante está compuesto de la siguiente jerarquía de archivos:

README.txt
Descripción de los archivos que conforman el corpus de cyberbullying.

cyberbullying_corpus.txt
Archivo de transcripciones de los diálogos con presencia de cyberbullying. El encabezado muestra el nombre de cada campo que compone el archivo separado por una tabulación. Los campos son: *id_instance* que es el identificador global de la línea de diálogo; *id_conversation* que es el identificador de cada conversación; *user_code* que es el código interno para los usuarios participantes en la conversación; y *transcription* que es la captura del diálogo.

global_classes.txt
Archivo que muestra una categorización de los diálogos por conversación. El encabezado del archivo muestra cada campo que lo compone, estos campos son separados por una tabulación. Los campos son: *id_conversation* que es el identificador de cada conversación; *id_bullying_level_code* que es el código del nivel de la agresión, sus valores son de [1-5]; *bullying_level que* es nivel de la agresión, puede tomar los siguientes valores [Agresión muy baja, Agresión baja, Agresión media, Agresión alta, Agresión muy alta]; *num_users* que es el número de participantes en la conversación; *num_male_users* que es el número de participantes del sexo masculino que intervienen en la conversación; *num_female_users* que es el número de participantes del sexo femenino que intervienen en la conversación; *topic_conversation* que es el tópico relacionado a la conversación, puede tomar los siguientes valores [Escolar, Recreativo, Familiar, Sociedad, Otro]; *agression_type* que es el tipo de agresión que se presenta en la conversación, puede tomar los siguientes valores [Rumores,

Amenazas, Comentarios, Exclusión, Compartir información confidencial, Otro]; *cyberbullying_category* que es la categoría del bullying identificado, puede tomar los siguientes valores [Racial, Sexual, Religioso, Apariencia, Desempeño académico, Nivel social, Otro]; e *interaction_type* que es el tipo de interacción encontrada entre los participantes en la conversación, puede tomar los siguientes valores [1-1, 1-N, N-1, N-M] que corresponden a [Uno a uno, Uno a muchos, Muchos a uno, Muchos a muchos] respectivamente.

post_classes.txt
Archivo que muestra una categorización por cada diálogo en las conversaciones. El encabezado del archivo muestra cada campo que lo compone, estos campos son separados por una tabulación. Los campos son: *id_instance* que es el identificador global de la línea de diálogo; *id_conversation* que es el identificador de cada conversación; *id_tone_code que* es el código del tono en el diálogo, sus valores son de [0-5]; *tone* que es el tono utilizado en el diálogo, puede tomar los siguientes valores [Sin agresión, Agresión muy baja, Agresión baja, Agresión media, Agresión alta, Agresión muy alta]; y *detonator_dialogue* que es el diálogo que inicia o detona la agresión en la conversación, puede tomar los valores de [0,1].

images/img{id_instance}.png
En esta carpeta se concentran las imágenes utilizadas en las conversaciones. Debido a la sensibilidad de los datos, personas presentes en las imágenes y contenido sexual han sido censurados, a excepción de personas públicas.

memes/meme{id_instance}.png
En esta carpeta se concentran las imágenes de tipo meme utilizadas en las conversaciones. Las imágenes memes están caracterizadas por presentar contenido visual acompañadas de frases.

additional/[meme|link]{id_instance}.txt
En esta carpeta se concentra información adicional. Los archivos 'meme{*id_instance*}.txt' contienen la transcripción del texto presentada en el meme. Los archivos 'link{*id_instance*}.txt' contienen vínculos externos a videos en YouTube.

Codificación de emoticonos, anonimato, imágenes y censurado

Aun cuando no se hace mención de nombre de personas completos, o se utiliza algún tipo de sobrenombres, se adoptó un esquema para anonimizar totalmente la identidad de los participantes, esto debido a la sensibilidad de los diálogos. Para cada nombre propio u referencia a participante se usó el siguiente código [@usuario{user_code}], y para externos a la conversación [@usuario_ext{#}]. Siguiendo esta misma definición para el contenido visual expresado en los diálogos se utilizó [img{id_instance}] y [meme{id_instance}]. Para el caso de emoticonos se utilizó [emoticono {adjetivo descriptivo}] donde adjetivo descriptivo puede tomar valores de llanto, sonrisa, enojado, popo, mano-arriba, etc. según el emoticono utilizado. Para el caso de múltiples emoticonos se usa un separador de ';'. Finalmente, en el caso de imágenes se procedió a incluir una marca de agua para censurar rostros o imágenes con contenido sexual o de agresión extrema.

4.2 Características del Corpus

En esta sección se describen las características del contenido del corpus lingüístico generado con el objetivo de resaltar su nivel de representatividad y calidad para utilizar en modelos de aprendizaje automático., Adicionalmente, se realiza un comparativo con dos corpus en español obtenidos de la red social Twitter (Garnacho, 2015), fuente comúnmente utilizada en estudios previos.

4.2.1 Características del corpus generado

El corpus generado presenta un contenido muy diverso respecto a la interacción entre los participantes, el tipo de conversaciones, y el nivel de agresión de estas. Existen características del corpus que lo hacen especialmente complejo. Complementario a su contenido inapropiado, se observa que los estudiantes acostumbran a comunicarse entre pares con palabras acortadas, escribir palabras modificadas intencionalmente, y a su vez tener poco cuidado con su ortografía. Algunos datos cuantitativos identificados en el corpus muestran que contiene 3,114 textos con un total de 19,587 palabras, de las cuales 3,802 son términos diferentes, y 2,516 términos aparecen solo una vez (frecuencia uno).

Del total de palabras con frecuencia 1:

1) 20% están mal escritas. Algunas de estas palabras muestran posibles errores no intencionales (ej. edsa, dl, esres, despernsa, abuerlo).
2) 30% de las palabras son modificadas intencionalmente. Por ejemplo, yisus, oie, okey, kha.
3) 39% son palabras poco usadas por los jóvenes, (ej. olmeca, debate).
4) 11% son no identificables (ej. waifu, boku, areglue).

Adicionalmente, se identificó un total de 379 emoticonos y otras combinaciones de caracteres para definir emociones (ej. :D, :C, xD, Dx, ;), :(, :'(, ;D).

Se identificó que los tópicos que se tratan con mayor frecuencia en las conversaciones son los relacionados con sociedad y recreación, con un 30% y 27% respectivamente, siendo el tema escolar un tema con un porcentaje significativo con el 17%; como se muestra en la gráfica de la Figura 6.

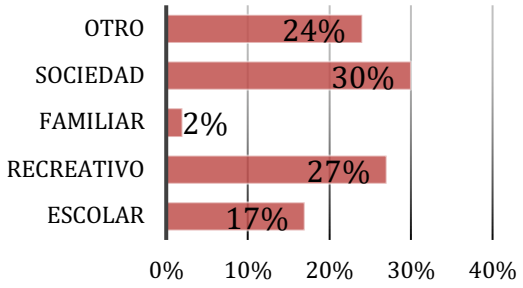

Figura 6. Tópicos presentes en las conversaciones del corpus.

La gráfica en la Figura 7 muestra los datos de las categorías de cyberbullying obtenidos del corpus. La categoría de nivel social obtuvo un 45%, destacando ampliamente sobre las otras categorías evaluadas por las psicólogas expertas en el área de cyberbullying. Lo anterior nos muestra que el nivel social o socioeconómico es un desencadenante de agresiones de cyberbullying.

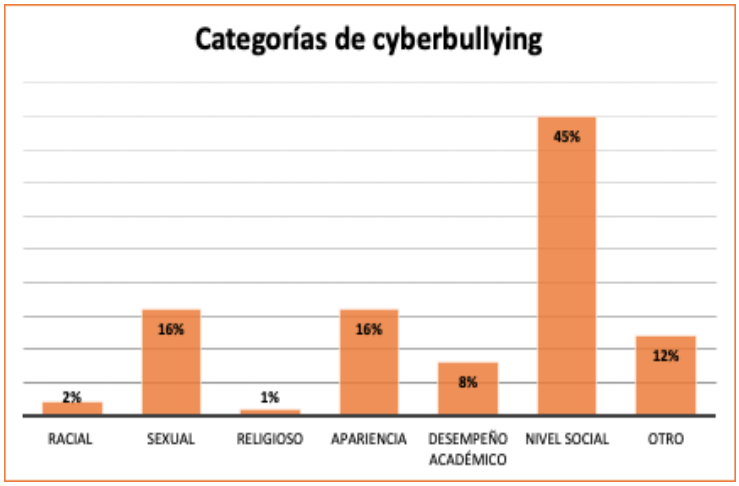

Figura 7. Categorías de cyberbullying.

Respecto al tipo de agresión realizada, en las conversaciones del corpus se encuentra que el acoso se realiza principalmente mediante comentarios, con un 61% de incidencia (ver Tabla 10).

Tabla 10. Tipo de agresión identificada en las conversaciones.

Tipo de Agresión	Porcentaje
Rumores	12%
Amenazas	6%
Comentarios	61%
Exclusión	12%
Compartir información confidencial	5%
Otro	4%

Las conversaciones en el corpus están conformadas en promedio por grupos de estudiantes de entre 2 y 3 personas. Pero las agresiones en las conversaciones se realizan principalmente en interacciones directas de 1 a 1 (72%), seguido por interacciones de muchos a 1 en un 19% (ver gráfica en la Figura 8).

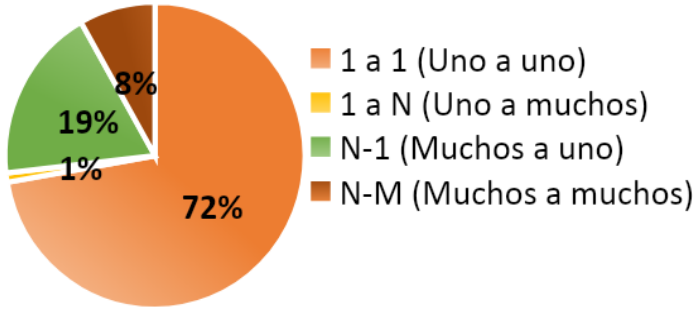

Figura 8. Interacción entre los participantes de las conversaciones.

Aun cuando se observan palabras y frases fuertes y/o altisonantes, con base en el criterio de las psicólogas etiquetadoras, el nivel de agresión en promedio fue considerado en el rango de bajo-medio para el 84% de las conversaciones del corpus (ver Figura 9).

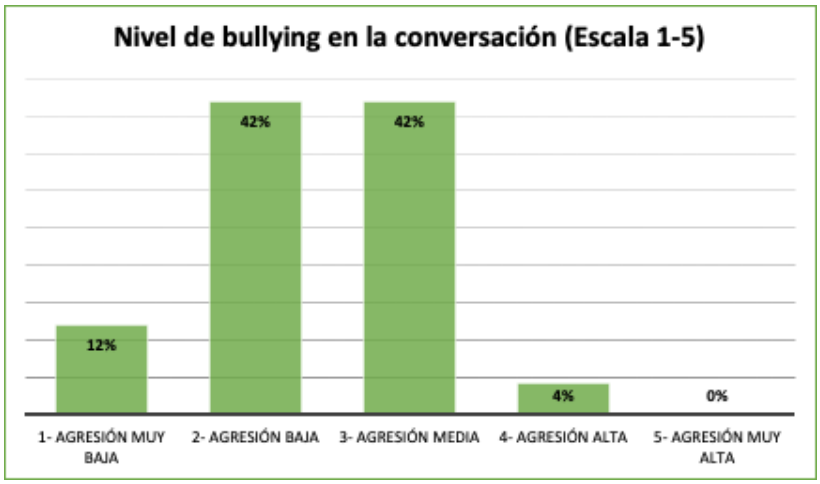

Figura 9. Nivel de bullying detectado en las conversaciones del corpus.

El análisis de los enunciados detonadores demostró que el 1% son comentarios que no presentan alguna agresión evidente, el 20% agresiones muy bajas y el 47% agresiones bajas. Los comentarios detonadores etiquetados con agresión media se presentan en un 30%. Solo el 2% de los enunciados detonadores es considerado con presencia de agresión alta (ver Figura 10).

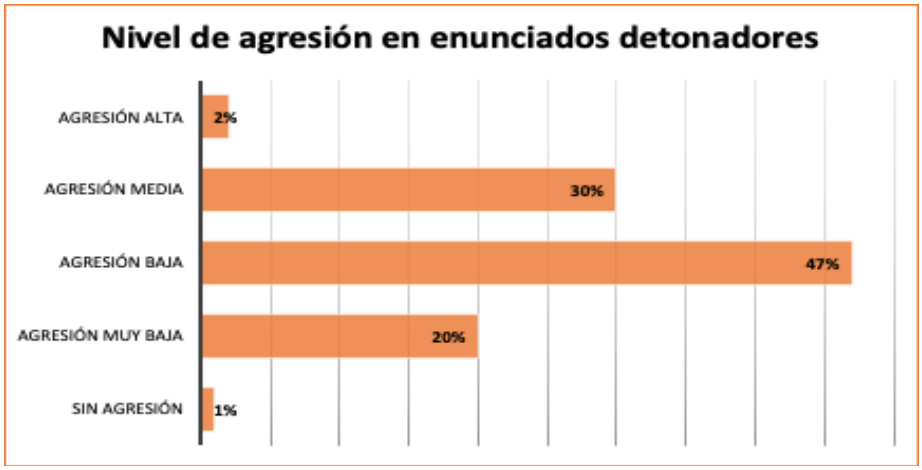

Figura 10. Nivel de agresión detectado en los enunciadores detonadores.

4.3 Análisis comparativo con otros corpus

Los diálogos obtenidos de las redes sociales que permiten una interacción mediante grupos privados, espacios utilizados principalmente por estudiantes de los niveles de estudio objetivo, presentan características no identificadas en corpus obtenidos de redes sociales públicas. Lo anterior genera implicaciones en el desempeño de los modelos predictivos generados.

Con el objetivo de analizar las características del corpus creado, se realizó un comparativo con dos corpus en español (multinacional) obtenidos de la red social Twitter, el primero etiquetado con comentarios positivos (Twitter-P) y el segundo con comentarios negativos (Twitter-N) (Garnacho, 2015). En este análisis, primeramente, se comparó el comportamiento de frecuencias entre palabras. Ya que la existencia de palabras poco frecuentes o con mucha frecuencia tienden a influir significativamente en el desempeño de los modelos de clasificación, al no ser elementos de referencia para identificar patrones de ocurrencia.

Una segunda comparación se centró en identificar el número de palabras reales del idioma español incluidas en los corpus. Para realizar este comparativo se utilizó un diccionario en español consistente de 80,328 palabras (Arce, 2014), considerado una referencia representativa, ya que la Real Academia Española (RAE) incluye 88,000 palabras en el idioma español.

Finalmente, se hizo un comparativo entre los corpus con el uso de palabras frecuentes del español mexicano. Para este último aspecto se utilizó un conjunto de datos consistente de las 1,000 palabras más utilizadas en el idioma español mexicano, elaborado por (Varela-Barraza et al., 2013). En la Tabla 11 se presentan los resultados obtenidos.

Tabla 11. Características de corpus provenientes de grupos privados y públicos.

Fuente	Número Textos	Número Palabras	Palabras Únicas	Palabras de Diccionario	Palabras Frecuentes	Etiquetado
Propio	3,112	17,827	**66%**	**36%**	**39%**	Manual
Twitter-P	55,360	42,323	62%	20%	12%	Automático
Twitter-N	122,216	63,772	60%	17%	7%	Automático

Aun cuando el tamaño del corpus creado para esta investigación es más pequeño que el utilizado en los otros dos conjuntos de datos, las características en su contenido son consistentes con respecto a la aparición de palabras únicas (frecuencia uno). Esto indica una tendencia a encontrar un alto número de palabras mal escritas, modismos o con errores ortográficos; independientemente de la red social de origen.

Inesperadamente, se observa una ligera superioridad en el corpus creado con respecto al número de palabras utilizadas pertenecientes al diccionario del idioma español de referencia. Se observa el comportamiento de que a mayor tamaño del corpus y mayor diversidad de procedencia de los usuarios es menor el porcentaje de palabras de diccionario. Finalmente, la característica que mayormente determina la representatividad del corpus generado, comparado con el de las otras dos fuentes, es sobre el uso de palabras frecuentes mexicanas. En este caso, utilizando el conjunto de palabras de uso común en el idioma español mexicano, se observa en el corpus propio un empleo elevado de estas palabras, inclusive un poco superior al porcentaje de palabras del diccionario.

En los corpus provenientes de Twitter, su contenido es generado por hispanoparlantes oriundos de diferentes países, implicando el uso de palabras y modismos distintos, palabras con significados diferentes entre una persona y otra, entre otros. La cantidad de palabras con frecuencia uno, así como la diversidad en palabras y modismos contenidos en un corpus multinacional, implica menor representatividad de un sector de la población y una mayor complejidad para el proceso de aprendizaje automático.

Con respecto a la calidad del corpus propuesto, aspecto considerado de mayor relevancia por la presente investigación, destacamos la presencia de tres características fundamentales, enfatizadas en la revisión de literatura sobre cyberbullying realizada por (Rosa et al., 2019; Elsafoury et al., 2021): 1) red social de procedencia, 2) balance del corpus, y 3) el etiquetado por expertos.

4.4 Análisis de Desempeño de Modelos de Clasificación

Con el objetivo de evaluar la representatividad del corpus generado para crear modelos de clasificación con desempeño aceptable, se llevó a cabo la implementación de modelos usando 6 de los algoritmos mayormente reportados en la literatura en atención a este problema.

4.4.1 Algoritmos de aprendizaje automático utilizados

Inicialmente, se consideraron los siguientes cuatro algoritmos de clasificación, tradicionalmente utilizados en aprendizaje automático (implementados mediante la librería scikit-learn):

1) Regresión Logística, método utilizado para predecir una clase binaria basándose en múltiples variables predictoras, con la capacidad de manipular matrices dispersas, propias de los vectores de texto. En este caso, se mantuvieron los valores por omisión propuestos en la librería, usando un máximo de 100 iteraciones para converger.

2) Máquinas de Vectores de Soporte, es un algoritmo de aprendizaje automático que se utiliza para clasificación y regresión de espacios no lineales que tiene la capacidad de separar un conjunto de observaciones en dos clases mediante un hiperplano de separación. Considerando que se cuenta con un corpus de tamaño pequeño, se optó por utilizar la implementación basada en libSVM con kernel lineal.

3) Clasificador Bayesiano Ingenuo, es un método de clasificación no lineal sencillo de implementar, pero con resultados satisfactorios en NLP. Particularmente, se implementó el clasificador NB multinomial, apropiado para la clasificación al tratar características discretas, útil para la categorización de textos.

4) Bosques Aleatorios, es referido como un método de tipo ensamble que busca incrementar el potencial del algoritmo de Árboles de Decisión, aumentando su desempeño y controlando el sobreajuste durante el entrenamiento. En la generación del modelo se usaron 1000 estimadores por grupo de árboles (bosques) y una profundidad máxima de árboles sin definir (hasta que todas las hojas fueran puras).

Adicionalmente, se crearon dos modelos de clasificación basados en aprendizaje profundo, la implementación se realizó utilizando las librerías TensorFlow y Keras:

1) Redes Neuronales Recurrentes (RNN, por sus siglas en inglés), es un tipo de red neuronal altamente utilizada en NLP ya que permite el procesamiento de datos con una naturaleza secuencial (Karpathy & Fei-Fei, 2015). Mediante un esquema de memoria a corto plazo, las RNN ponderan la importancia del orden de aparición de las características (palabras).

   En esta investigación se implementó un modelo de 3 capas usando la clase *Sequential* de Keras. Primero, se incluyó una capa de incrustación que suministra los enunciados mediante un vector de palabras, con un tamaño de vocabulario de 2,000 palabras y una dimensión de salida de 100 neuronas. Posteriormente, se implementó una capa de memoria de corto plazo extendida

(LSTM, por sus siglas en inglés) con 100 neuronas intermedias y una función de activación tangente hiperbólica; útil en la clasificación de texto al ser efectiva para aprender secuencias y atender el problema de desvanecimiento de gradiente (Pascanu et al., 2013). Finalmente, se usó una capa densa de predicción con una sola neurona que representa una salida binaria. El modelo se generó ejecutando 10 épocas de entrenamiento con una manipulación de lotes de 30 sentencias. En el experimento se observo que al usar un mayor número de épocas (10, 50, 100), no mejoraba el desempeño de clasificación del modelo, esto debido al tamaño reducido del corpus. Sustentando que el incremento en el número de épocas tiene sentido cuando se cuenta con un corpus de tamaño superior.

2) Redes Neuronales Convolucionales (CNN, por sus siglas en inglés), son usadas típicamente en problemas de clasificación de imágenes basándose en la identificación de características o rasgos principales (Karpathy & Fei-Fei, 2015). Esta modalidad de red neuronal también ha sido experimentada con éxito en el dominio de NLP (Kim, 2014). El entrenamiento se realiza al convertir una cadena de texto a una matriz bidimensional, la cual recibe el proceso de convolución similar al utilizado con imágenes en escala de grises.

   Para la implementación de CNN, primeramente, se procedió a *tokenizar* los enunciados en vectores numéricos ajustados a una misma longitud, donde cada valor representa a una palabra diferente en el vocabulario del corpus. En total se implementaron 7 capas: 1) la capa de incrustación definida en un espacio vectorial de 200 valores asignados a cada palabra; 2) dos capas que implementan una convolución unidimensional mediante 100 filtros de dos palabras (*bigramas*) y tres palabras (trigramas), utilizando la función de activación rectificadora lineal unitaria (*ReLU* por sus siglas en inglés); 3) una capa global de *max-pooling* unidimensional llamada después de cada capa de convolución; y 4) una capa densa (oculta) con 256 neuronas y la función de activación *ReLU*; 5) una capa de *dropout* al 20% para reducir el sobreajuste; y 6) la capa densa para realizar el proceso de clasificación binario con la función de activación sigmoide. Finalmente, el modelo generado logro un adecuado desempeño ejecutando 5 iteraciones de entrenamiento con una manipulación por lotes de 32 sentencias; nuevamente, el tamaño del corpus permitió lograr un desempeño adecuado con un número reducido de iteraciones de entrenamiento; Compilando con la función de pérdida de entropía cruzada binaria y el optimizador Adam (Estimación Adaptativa de Momentos).

4.4.2 Preprocesado de los datos

En investigaciones como la presentada en esta tesis los recursos de texto provienen de redes sociales. Estos textos tienen características particulares originadas por el estilo peculiar de escritura de los usuarios, usualmente con presencia de errores ortográficos, abreviaciones, modismos, entre otros estilos de escritura. Categorizando estos textos como de baja calidad

lingüística. Es por ello que se requiere de un proceso de limpieza que permita mejorar la calidad del texto. A este proceso se le conoce como preprocesamiento.

En esta fase se recibe como entrada un conjunto de datos de texto en lenguaje natural y genera como salida un texto optimizado, conservando la coherencia del texto original y con una mejor calidad. Un conjunto de datos preprocesados donde se han eliminado datos incompletos, ruidos, errores, entre otros, permitirá una mayor comprensión de la información al momento de modelar su contenido (Salama, Kader & Abdelwahab, 2021).

En esta fase, se realizó el conjunto de transformaciones comúnmente utilizadas en el tratamiento de texto, las cuales consistieron en: 1) la eliminación de dígitos y caracteres especiales (incluyendo acentos), 2) reducción de énfasis, 3) normalización de expresiones de risa y otras expresiones especiales, 4) normalización a singular y minúsculas, 5) eliminación de *stopwords* (palabras vacías o con poco significado), y 6) obtención de la raíz de las palabras (*stemming*).

4.4.3 Implementación de algoritmos de clasificación

Una vez construido el corpus, considerando que su tamaño es significativamente menor a los utilizados en investigaciones previas, se evaluó su representatividad para construir modelos de clasificación para identificar expresiones de cyberbullying con un nivel de precisión aceptable. Se procedió a generar los modelos de clasificación usando los principales algoritmos de aprendizaje automático reportados en la literatura, descritos en la sección previa. Se definió y utilizó el mismo conjunto de datos de entrenamiento (70%) y prueba (30%) para generar y evaluar todos los modelos; los cuales fueron obtenidos de forma aleatoria del corpus original. Como una característica importante, el total de textos etiquetados con presencia de bullying fue del 65% para el conjunto de datos de entrenamiento y del 56% en los datos de prueba, lo cual muestra una proporción aceptable entre el número de observaciones de cada clase.

El desempeño de los modelos para identificar enunciados con presencia de bullying se midió considerando las siguientes métricas de evaluación: exactitud, precisión, sensibilidad y Valor-F. En la Tabla 12 se presenta el desempeño obtenido por cada uno de los algoritmos utilizados. Considerando como referencia comparativa la métrica de Valor-F, la más utilizada en estudios previos, podemos observar que el clasificador Bayesiano Ingenuo fue el modelo con el mejor desempeño.

Tabla 12. Desempeño de los algoritmos de clasificación.

Algoritmo	Exactitud	Precisión	Sensibilidad	Valor-F
LR	0.766	0.794	0.924	0.854
SVM	0.717	**0.818**	0.794	0.806
RF	0.754	0.761	**0.974**	0.854
NB	**0.773**	0.782	0.961	**0.862**
RNN	0.747	0.775	0.929	0.845
CNN	0.705	0.814	0.800	0.807

NB superó ligeramente los modelos generados con variantes de redes neuronales. Este desempeño inferior de los modelos de aprendizaje profundo se atribuye al tamaño reducido del corpus y la alta cantidad de palabras con frecuencia uno. Así mismo, aun cuando SVM es un algoritmo que tiende a ser eficaz en espacios de grandes dimensiones, y se esperaba un mejor desempeño, aparentemente su precisión se vio afectada por contar con un corpus con una gran cantidad de características (palabras) y un reducido número de muestras. Sin embargo, se obtuvo un mejor desempeño que el reportado en estudios previos que utilizaron este mismo algoritmo (ver Tabla 13).

Tabla 13. Comparativo de desempeño de algoritmos de clasificación basado en la métrica de Valor-F (V-F) y exactitud (EXAC).

Fuente	Medida	LR	SVM	RF	NB	RNN	CNN	PROP
Propio	V-F	0.854	**0.806**	0.854	**0.862**	**0.845**	0.807	
Balakrishnan et al., 2019	V-F			**0.929**				
Rosa et al., 2019	V-F	0.740	0.750	0.650				
Banerjee et al., 2019	EXAC						**0.939**	
Chelmis et al., 2019	V-F	**0.868**						
Kumar et al., 2019	EXAC			0.740	0.644			
Samghabadi et al., 2020	V-F		0.760					
Wang et al., 2020	V-F							0.860
Fortunatus et al., 2020	V-F							**0.866**

* PROP (método de clasificación con mejor desempeño propuesto por los autores)

Se observa que, comparando con los resultados reportados por la literatura, al utilizar RF se obtuvo un desempeño superior al logrado con SVM. RF está basado en un método sencillo, pero bastante funcional para aprender relaciones complejas altamente no lineales, similares a las generadas por el alto número de palabras únicas en el corpus. Al utilizar RF se logró consistencia con lo reportado en estudios previos, observando una precisión superior a la reportada en (Rosa et al., 2019) y ligeramente inferior a la reportada en (Balakrishnan et al., 2019). Es importante mencionar que el menor desempeño obtenido en (Rosa et al., 2019) puede derivar de haber utilizado un corpus de 2,999 tweets, similar al generado en esta investigación, pero mucho menor a lo utilizado en otros estudios.

Al utilizar NB, un modelado probabilístico ampliamente utilizado en tareas de clasificación de textos, se logró el mejor desempeño con un Valor-F de 0.86. Lo anterior, aun

cuando el corpus utilizado contiene una alta cantidad de palabras con frecuencia uno, consideradas como características raras o irrelevantes que causan mayor problema en el desempeño de este algoritmo. Se asume que el equilibrio de observaciones entre las clases, y la capacidad del algoritmo de manejar independencia entre variables predictoras, permitió un buen nivel de clasificación de este modelo.

Aun cuando en la literatura se reporta un desempeño prometedor al utilizar aprendizaje profundo, la precisión obtenida en la presente investigación fue aceptable pero no superior a lo logrado con NB. Sin embargo, la RNN logró un desempeño muy cercano. Adicionalmente, aún cuando las redes neuronales convolucionales (CNN) son mayormente utilizadas para clasificar imágenes, en la literatura se reportan buenos resultados en el ámbito del procesamiento de lenguaje natural (Banerjee et al., 2019). En este experimento se observa la influencia de palabras que aparecen una sola vez, haciendo muy difícil que este tipo de red neuronal les pueda dar un significado y un contexto, para que pueda entender cuando se utilizan.

Finalmente, se realizó un experimento preliminar considerando la influencia de múltiples características para el entrenamiento del modelo de clasificación con mejor desempeño (NB). Inicialmente, seleccionamos cinco características: 1) *Mensaje_detonante* (mensaje que detona una situación de acoso dentro de una conversación, etiquetado con 1 (detonador) o 0 (no-detonador)); 2) *Frecuencia_uno* (valor numérico que indica la cantidad de palabras únicas dentro del mensaje); 3) *Participantes* (valor numérico que indica la cantidad de participantes en esa conversación); 4) *Participantes_femeninos* (valor numérico que indica la cantidad de mujeres participando en esa conversación); y 5) *Participantes_masculinos* (valor numérico que indica la cantidad de hombres participando en esa conversación).

Mediante una prueba Chi-cuadrada, usando la clase *SelectKBest* de scikit-learn, las siguientes tres características resultaron como mejores predictoras de la variable dependiente: *Mensaje_detonante*, *Frecuencia_uno* y *Participantes*. Sin embargo, los resultados obtenidos en el experimento indicaron un incremento mínimo en el desempeño del modelo usando NB (ver Tabla 14). Se considera como trabajo futuro realizar un estudio más detallado sobre el análisis multivariable para la identificación de cyberbullying considerando lo estipulado en (Elsafoury et al., 2021).

Tabla 14. Análisis multivariable usando el algoritmo de clasificación NB.

Característica(s)	Valor-F
Característica de texto usada en experimento original (**CT**)	0.862
CT + Mensaje_detonante	0.864
CT + Frecuencia_uno	0.864
CT + Participantes	0.861
CT + Mensaje_detonante + Frecuencia_uno	**0.866**
CT + Mensaje_detonante + Frecuencia_uno + Participantes	0.865

Capítulo 5

Ambiente de evaluación

5.1 Diseño del ambiente de evaluación

Con el objetivo de evaluar el modelo de clasificación probabilístico Naive Bayes univariante en un ambiente controlado con usuarios reales, algoritmo que logró el mejor desempeño con un Valor-F de 0.86 en la tarea de clasificación de textos, se buscó corroborar su funcionamiento en un ambiente de interacción social categorizando en tiempo real mensajes realizados en conversaciones expuestas a situaciones de cyberbullying. Se desarrolló MyBook Blog, un sistema Web con una temática tipo blog que permite crear un escenario de una conversación real donde cada comentario sería evaluado con el modelo antes mencionado.

5.1.1 Requerimientos funcionales

Previo a la implementación del ambiente de evaluación se procedió a definir las características o requerimientos funcionales que dicho entorno debiera cumplir, considerando cinco aspectos principales:

1) *ambiente de interacción*, generando un entorno que permitiera el libre dialogo entre participantes;
2) *tópico detonador*, poder detonar situaciones de conflicto de opinión durante una conversación;
3) *notificación de bullying*, mostrar en tiempo real, a los participantes en la conversación, el etiquetado de los mensajes intercambiados realizado por el modelo de clasificación propuesto;
4) *evaluación de bullying*, permitir que los participantes en el dialogo evalúen el etiquetado del modelo; y
5) *registro de interacción*, llevar un registro de todo los mensajes e interacciones realizadas en una conversación.

Ambiente de interacción

Se optó por utilizar un entorno de comunicación social tipo blog dado sus características de interacción, siendo un espacio de intercambio de mensajes basado en Web donde se discuten o informan aspectos informales con un estilo de escritura libre. Particularmente, considerando las peculiaridades funcionales del ambiente requerido, el sistema se generó desde cero utilizando el lenguaje de desarrollo PHP, CSS, HTML y Java Script, utilizando el manejador de base de datos convencional MySQL.

Adicional a las funcionalidades comunes de un espacio tipo blog, donde un participante puede publicar noticias o comentarios, y estas son presentadas en orden cronológico, mostrando el último mensaje en la parte superior del historial, se procedió a implementar las funcionalidades adicionales requeridas para evaluar el modelo de clasificación propuesto.

Tópico detonador

Se planeo contar con un ambiente de interacción donde los participantes seleccionaran una sala de su interés, con base en la discusión de algún tópico de vanguardia y polémico para jóvenes. Los temas propuestos atienden tópicos como "*Si o no a la vacuna para COVID-19*" y "*Feminismo*." Adicionalmente, una vez que el participante entra a la sala se incluyen preguntas que fomenten el inicio del dialogo, tales como "*Estas de acuerdo en que se vacune a toda la población*." Finalmente, considerando mantener un grupo reducido y controlado de participantes, similar al número de identificado en grupos cerrados de amigos, se solicitó el poder configurar la cantidad de participantes en cada sala.

Notificación de bullying

Se estableció la necesidad de que el modelo de clasificación etiquetara cada una de las publicaciones en la conversación. Así como, mediante un mensaje y código de colores, se le mostrara a cada participante, emisor y receptores, si un mensaje fue categorizado con presencia o ausencia de bullying. Se considero mostrar tanto categorizaciones positivas (presencia de bullying) como negativas (ausencia de bullying), con la finalidad de identificar falsos positivos y falsos negativos.

Evaluación de bullying

Complementando la notificación de clasificación de los mensajes intercambiados, con presencia o ausencia de bullying, se definió que cada una de las categorizaciones de comentarios en el blog incluyera la opción para que el usuario indicara si estaba de acuerdo o no con la clasificación realizada por el modelo.

Adicional a poder enviar comentarios, la evaluación de cada uno de los mensajes la podrían realizar solamente los receptores, no dando opción al emisor de indicar si estaba de acuerdo con la clasificación del modelo. En dicha evaluación se estableció que debía ser realizada de manera rápida, a través de elementos conocidos (botones), y que permitiera la opción de modificar una evaluación (para situaciones donde el participante reflexione y cambie de opinión).

Registro de interacción

Como requisito fundamental para poder evaluar el desempeño del modelo para clasificar comentarios con presencia o ausencia de bullying, se definió necesario registrar la clasificación realizada por el modelo, así como la valoración de la categorización por parte de los participantes en la conversación.

5.1.2 Interfaz gráfica del sistema MyBook Blog

Inicialmente el sistema muestra una vista de registro donde el alumno debe de crear una cuenta y posteriormente iniciar sesión en el sistema para proceder con la selección de una sala de conversación (ver Figura 11).

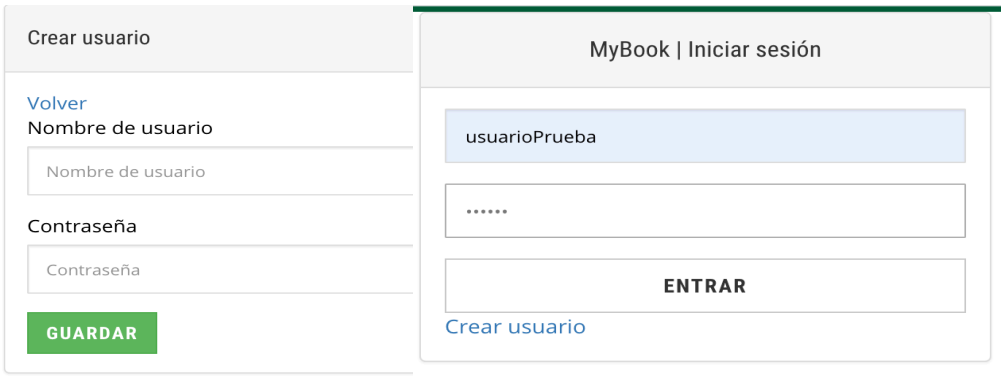

Figura 11. Vistas de registro e inicio de sesión del sistema MyBook Blog.

En el sistema se muestran las salas de conversación disponibles, indicando el tema central de conversación, que permite a los estudiantes seleccionar una conversación de interés donde dialogarán con otros participantes (ver Figura 12). Los temas seleccionados para las salas son temas controversiales o polémicos donde se propicia el intercambio de opiniones y se asume abre las puertas para desencadenar una situación de cyberbullying.

Figura 12. Interfaz en el sistema para la selección de sala de conversación.

Una vez que los estudiantes han seleccionado un tema de conversación el sistema les mostrará la vista de la sala que contiene el título o tema, la descripción de lo que se está solicitando inicialmente opinar para dar inicio al intercambio de ideas, y los comentarios realizados por los demás participantes sobre ese tema (ver Figura 13).

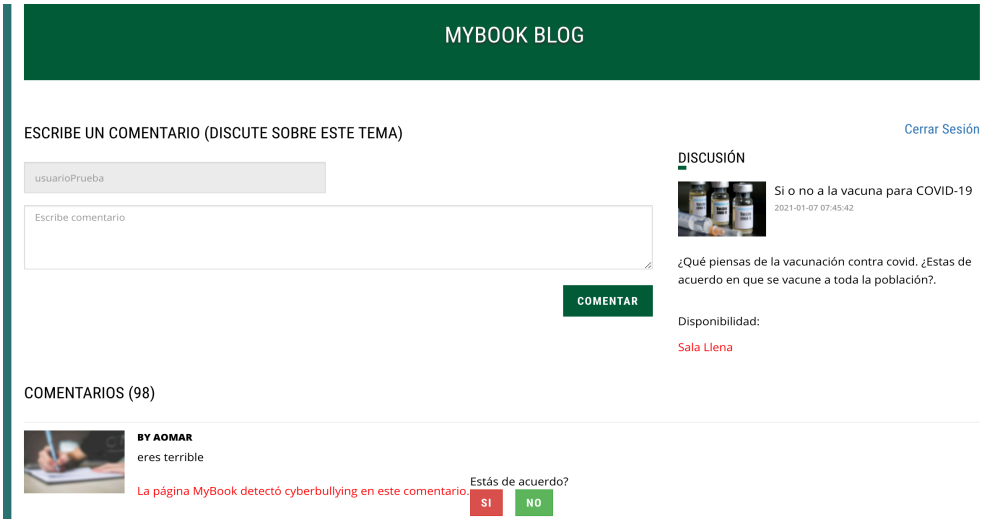

Figura 13. Vista de una sala de conversación.

Cada comentario realizado por los participantes en la sala de conversación es evaluado con el modelo generado con NB para determinar si es considerado con presencia de cyberbullying o no. Adicionalmente, por cada comentario realizado por los usuarios se muestra una opción a votación en aquellos que el algoritmo ha detectado con presencia de cyberbullying. Esto permitirá corroborar la efectividad del algoritmo en cuanto a las clasificaciones catalogadas con presencia de cyberbullying, ya sea clasificadas correcta o erróneamente (falsos positivos). Esta votación se realiza sobre los comentarios realizados por los demás participantes de la conversación, evitando la votación de los comentarios propios, lo anterior con el propósito de mantener la objetividad de la votación.

De modo complementario a la vista de usuario, se creó una vista de administrador que permite crear nuevas salas de conversaciones (ver Figura 14), así como administrar los comentarios creados bajo cada sala y revisar la clasificación asignada por el modelo (ver Figura 15).

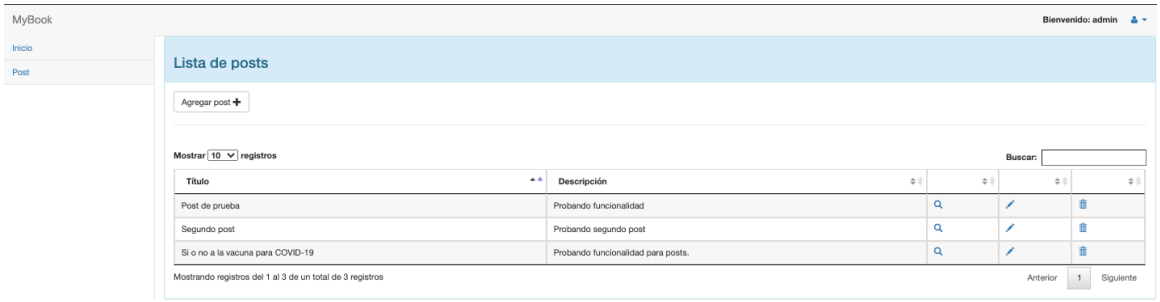

Figura 14. Vista del administrador en el sistema. Aquí se crean nuevas salas de conversación y se administran las ya creadas.

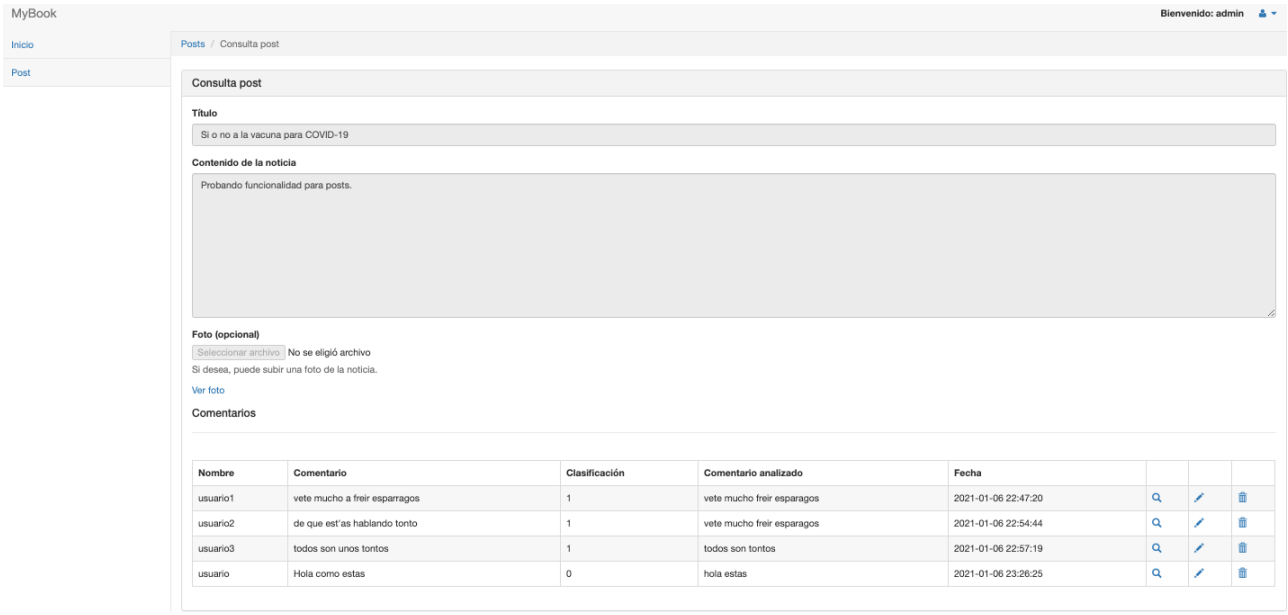

Figura 15. Vista del administrador en el que permite administrar los comentarios creados en una sala en específico.

5.2 Arquitectura del sistema propuesto

La arquitectura propuesta para el sistema de evaluación se muestra en la Figura 16. En general, la arquitectura consta de un equipo de cómputo que hospeda la aplicación implementada (MyBook Blog), el servidor web Apache, el manejador de base de datos MySQL y el intérprete del lenguaje PHP. El flujo de procesamiento al usar el sistema se describe a continuación.

Inicialmente, el sistema recibe el comentario escrito por el usuario y procede a guardarlo en la base datos, paralelamente el comentario pasa por una etapa de preprocesamiento para "limpiarlo", utilizando los mismos algoritmos de preprocesado que se usó al limpiar los datos de entrenamiento del modelo.

Durante el proceso de limpieza se procede a eliminar o convertir cualquier carácter especial que se utilice de alguna manera. Por ejemplo, se realizan transformaciones para formar un emoticono como :), :(, xD, ;), entre otros. Además, se eliminan aspectos del lenguaje como signos de puntuación y acentuación, eliminación de caracteres repetidos, eliminación de dígitos, reducción de énfasis, normalización de expresiones de risa y otras expresiones especiales, normalización a singular y minúsculas, eliminación de *stopwords*, y obtención de la raíz de las palabras (*stemming*).

Una vez que un comentario pasa por el proceso de limpieza y obtenemos su versión corregida, el sistema lo guarda en la base de datos con la distinción de que es el comentario preprocesado. Posteriormente, utilizando el modelo creado con NB, se procede a realizar la clasificación para determinar si es etiquetado con contenido o no de cyberbulling.

Finalmente, como parte del proceso implementado, en el sistema se muestra la clasificación dada por el algoritmo, además de una leyenda con dos botones. Estos botones tienen la finalidad de que los usuarios puedan votar si están de acuerdo o no con la clasificación realizada por el modelo, y esta votación se guarda en la base de datos en conjunto con la clasificación obtenida. Dichos botones solo se muestran sobre los comentarios realizados por los demás usuarios y no sobre los propios.

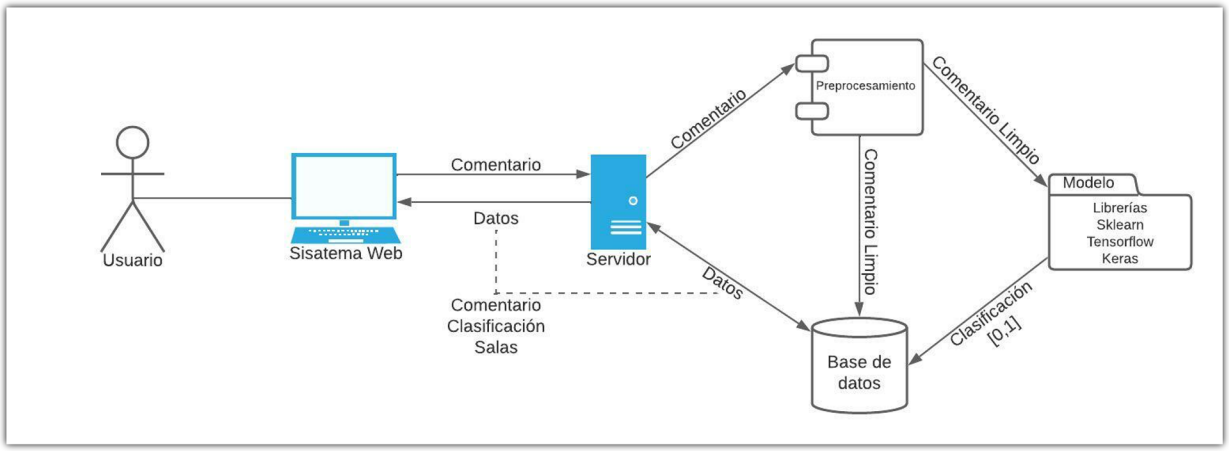

Figura 16. Arquitectura del ambiente evaluación.

5.3 Evaluación de MyBook Blog (Sistema Web) implementando el modelo de clasificación Naive Bayes

Para la evaluación de funcionalidad del sistema se establecieron ciertas especificaciones que propiciarán un ambiente sugerente para el desarrollo de una conversación que pudiera desencadenar una situación de cyberbullying. Esta evaluación de funcionalidad se generó en un ambiente controlado, el cual contó con la participación de un usuario comodín que se adaptaría al flujo de la conversación y fomentaría la participación en el mismo, recordando que el objetivo principal era identificar mensajes con presencia de bullying. Este proceso de evaluación consistió de 4 fases: 1) selección de participantes, 2) selección de temas controversiales, 3) registro de participantes, y 4) monitoreo de uso del sistema.

Fase 1

En una primera fase de evaluación se seleccionó a un grupo de estudiantes que cumplieran con el perfil de la población objetivo: estudiantes de nivel medio superior o superior, en este caso se logró la colaboración de alumnos pertenecientes a la Universidad Autónoma de Baja California, estudiantes de distinto sexo, estudiantes que no se encontraban en el mismo espacio físico, y finalmente que pertenecían a distintos grupos académicos (no se conocían entre ellos).

Finalmente, se seleccionó a un pequeño grupo de estudiantes para evaluar el sistema consistiendo en 2 participantes del género femenino y 1 participante masculino. Dado que esta

evaluación se desarrolló en un ambiente controlado se contó con un participante extra que se debía adaptar a la conversación y generar comentarios en dado caso que los estudiantes se mostraran poco participativos durante la actividad.

Fase 2

Durante la fase 2 de la evaluación se identificaron temas controversiales actuales que incitaran al diálogo, considerando el perfil y sexo de los participantes. Algunos de los temas considerados para la actividad estaban relacionados a experiencias sociales vigentes durante el periodo de evaluación. Los temas abarcaron desde intereses extracurriculares, hobbies, política, situación actual de COVID-19, asuntos socio políticos, de religión y racismo.

Finalmente, se optó por usar como tema central el feminismo para desarrollar la sala de conversación, dado que los participantes seleccionados eran de ambos sexos.

Fase 3

Una vez seleccionados los participantes y el tema de conversación, se creó la sala en el sistema MyBook con el objetivo de crear el escenario de control para probar la funcionalidad. A los estudiantes participantes se les proporcionó un manual electrónico donde se les guiaba sobre cómo realizar el proceso de registro en el sistema y seleccionar la sala de conversación. Después de seguir los pasos de registro y selección de la sala, se les mostró a los estudiantes el objetivo de la conversación y los comentarios realizados dentro de esa sala, así como el formulario para crear un nuevo comentario y evaluar las clasificaciones del modelo.

Fase 4

Finalmente, cuando los participantes continuaron con la actividad e iniciaron el intercambio de opiniones o discusión, se les fueron mostrando las opciones de votación por cada comentario realizado por los demás participantes de la sala, solicitándoles valorar la correcta clasificación realizada por el sistema.

Para finalizar con la actividad de evaluación se realizó un análisis estadístico de la información obtenida.

5.4 Resultados de la evaluación

Una vez concluido el experimento se analizaron los resultados obtenidos durante la evaluación del sistema, descritos en la Tabla 15. Podemos observar información sobre la cantidad de interacciones y el número de clasificaciones correctas e incorrectas, de acuerdo con lo que clasificó el modelo y opinaron los estudiantes.

Tabla 15. Clasificaciones realizadas por el modelo Naive Bayes en el ambiente de evaluación.

Sistema	Cyberbullying (8 clasificaciones)			No cyberbullying (18 clasificaciones)		
	TP	FP	NC	TN	FN	NC
Usuario 1	1	2	5	12	1	5
Usuario 2	2	5	1	8	3	7
Usuario 3	1	4	3	11	0	7
Revisor	5	3	0	17	1	0

*TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative), NC (No Comentó)

En la Tabla 15 se muestran los resultados obtenidos por el modelo Naive Bayes en el ambiente durante la evaluación en el escenario de control. De un total de 26 mensajes intercambiados, se observa que el sistema clasificó como Cyberbullying 8 de los comentarios realizados por los usuarios y 18 como comentarios sin presencia de cyberbullying. Dentro de estas categorizaciones el Usuario 1 detectó 2 falsos negativos para las clasificaciones de cyberbullying y 1 falso negativo para las clasificaciones realizadas como No-cyberbullying, adicionalmente dicho usuario se abstuvo de votar en 5 de los comentarios realizados en la conversación, es decir aquellos comentarios realizados por él mismo. El usuario 2 detectó 5 falsos positivos y 3 falsos negativos. Por último, el usuario 3 detectó 4 falsos positivos y ningún falso negativo.

Considerando los resultados de etiquetado del modelo con usuarios reales, presentados en la Tabla 15, con base en la evaluación de los participantes en la conversación, observamos un desempeño muy desfavorable del modelo. Tomando como base la métrica Valor-F observamos que con los tres usuarios participantes en el experimento el modelo obtuvo un Valor-F de entre 0.332 y 0.398 (ver Tabla 16). Con el objetivo de identificar las causas de este bajo desempeño un revisor adicional analizó cada una de las publicaciones realizadas por los participantes; miembro del grupo de investigadores participantes en el estudio y considerando las categorías de agresión mencionadas en el capítulo 4 (sección 4.1.3). Se contabilizó, a criterio del revisor adicional la categoría asignada por el modelo y la opinión vertida por los mismos participantes en la conversación. En el Apéndice C se incluyen los diálogos de esta conversación, la categorización asignada por el modelo a cada dialogo, la evaluación de la categorización por parte de los participantes en la conversación, y la evaluación de la categorización por parte del revisor adicional.

Tabla 16. Desempeño del modelo de clasificación NB.

Evaluador	Valor-F
Usuario 1	0.398
Usuario 2	0.332
Usuario 3	0.332
Revisor	0.833

En la Tabla 16 podemos observar que el Valor-F obtenido de la evaluación del revisor adicional da un resultado de 0.833, muy cercano al 0.862 que se logró en la fase de pruebas del modelo. Este resultado se logra al mantener una consistencia en las categorizaciones correctas de publicaciones sin presencia de bullying y al incremento de categorizaciones correctas de publicaciones con presencia de bullying. Además, el revisor adicional evaluó el total de publicaciones (26), y los participantes en la conversación solo revisaron 16 y 18.

Adicionalmente, considerando que la mayor discrepancia entre las evaluaciones de los participantes en la conversación y el revisor se dieron al valorar el desempeño del modelo para identificar comentarios con agresión, se identificó que en ocasiones si hubo coincidencia en la valoración y en otras no. Por ejemplo, todos estuvieron de acuerdo en que la siguiente publicación fue correctamente categorizada "*la mujer no sirve para nada, solo para criar hijos y cuidar la casa, que se vayan a cocinar mejor*". Sin embargo, en publicaciones como "*me parecen comentarios tontos, debemos de participar con sentido no haciendo comentarios nada mas por que si*" y "*no leiste las instrucciones del correo o que*" se encontraron discrepancias. Enfatizando que en la última de las publicaciones mencionadas se encontró que algunos participantes indicaron que la categorización del modelo fue correcta y otros estuvieron en desacuerdo.

Capítulo 6

Conclusiones

En este capítulo se presenta un análisis de las implicaciones del estudio en relación con las preguntas y objetivo de la investigación, para lo cual se establecen los métodos aplicados para dar respuesta y cumplimiento al respecto. Además, se describe el aporte al área de conocimiento y se proponen sugerencias para futuras investigaciones que se encaminen en estudiar el fenómeno de cyberbullying abordado en la presente investigación.

6.1 Sobre las preguntas de investigación

Con el objetivo de dar respuesta a la pregunta de investigación **¿Qué factores detonan y permiten identificar situaciones de cyberbullying en entornos de redes sociales?**, se realizó una revisión de la literatura que nos permitió identificar la categorización para diálogos de conversaciones realizados por la población objetivo. De acuerdo con (Kurniasih et al., 2020; Ansary, 2020; Berdugo-Gómez, 2020; Ruiz-Ramírez et al., 2020; Coob & Marín, 2020), los tópicos, categorías y formas de agresión derivan de lo expuesto en la literatura, dando como resultado la categorización de diálogos considerando el tópico de la conversación (Escolar, Recreativo, Familiar, Sociedad, Otro), la categoría de bullying (Racial, Sexual, Religioso, Apariencia, Desempeño académico, Nivel social, Otro), y el medio o la forma en que se realizaban los ataques (Rumores, Amenazas, Comentarios, Exclusión, Compartir información confidencial, Otro). La identificación de estos conceptos relacionados a las prácticas de bullying permitió generar el documento de clasificación presente en el Anexo B para evaluar los diálogos de los estudiantes participantes en la generación del conjunto de datos, que posteriormente utilizarían la terna de psicólogas para clasificar cada conversación y diálogo obtenido en el proceso de desarrollo de esta investigación.

En respuesta a la pregunta de investigación **¿Qué modelo de aprendizaje automático o aprendizaje profundo nos permite realizar una mejor clasificación de los factores detonantes de cyberbullying, en términos de precisión?**, se consideró en la fase de modelado la utilización de algoritmos propios de aprendizaje automático, destacando en desempeño el modelo basado en NB. Adicionalmente, se generaron modelos usando algoritmos de aprendizaje profundo. Los experimentos realizados permiten observar y ratificar al campo de estudio, que el tamaño del corpus influye negativamente en el desempeño de los modelos generados con técnicas de aprendizaje profundo. Logrando un mejor desempeño con algoritmos de aprendizaje automático tradicionales. Sin embargo, en general, los resultados obtenidos fueron superiores o muy cercanos a los reportados en estudios previos.

Finalmente, es importante enfatizar que las tareas de preprocesado consideradas en el estudio influyeron en el desempeño final de los modelos generados. Sin embargo, se resalta la existencia de un alto número de palabras únicas, considerando un área de investigación importante su reducción mediante la edición de palabras mal escritas, y la identificación y

normalización de términos de la jerga usada por jóvenes en las redes sociales. Así como, adicional al uso de los emoticonos, los datos multimodales del corpus deben ser considerados mas ampliamente en la generación de modelos como características que permitan extender el análisis multivariable presentado en la presente investigación.

Para evaluar la pertinencia de usar un corpus representativo de la audiencia objetivo, pero relativamente menor al utilizado en otras investigaciones, se generaron un conjunto de modelos para detectar situaciones de cyberbullying usando algoritmos tradicionales de aprendizaje automático y variantes de aprendizaje profundo. Se evaluó la eficacia de los modelos generados, obteniendo un desempeño similar, y en ocasiones superior, al reportado en estudios previos. Aun cuando el desempeño obtenido fue significativamente menor al de mejor desempeñó reportado por otros investigadores que utilizaron Redes Neuronales Convolucionales (Valor-F 0.939), mediante el Clasificador Bayesiano Ingenuo se obtuvo un Valor-F aceptable de 0.862. Con los resultados obtenidos se observa que el tamaño del corpus y la alta cantidad de palabras únicas (Frecuencia uno) influyen negativamente en el desempeño de los modelos generados con técnicas de aprendizaje profundo, logrando un mejor resultado con algoritmos de aprendizaje automático tradicionales.

6.2 Sobre los objetivos de investigación

Con relación al objetivo de la investigación que guió los procesos de el presente estudio: Crear un *modelo de clasificación*, robusto y escalable, es decir que pueda ser adaptado para incluir otros lenguajes o jergas y una mayor cantidad de datos a analizar, que permita identificar situaciones de cyberbullying en espacios virtuales, construido y validado con datos multimedia que estos mismos entornos generan.

El diseño de investigación aplicado para cumplir con el objetivo corresponde a un enfoque que se sustenta en la metodología para la implementación de modelos de aprendizaje automático propuesta en, la ***Machine Learning Pipeline (ML pipeline)***, que consiste en una serie de pasos ordenados que establecen la secuencia de trabajo de aprendizaje automático ((Elsafoury et al., 2021). Este diseño de la investigación permitió estudiar a profundidad el fenómeno de cyberbullying siguiendo la secuencia de la ML Pipeline.

En la lucha contra el fenómeno de cyberbullying, considerando la importancia social de atender esta problemática, diversas investigaciones se han enfocado en definir estrategias o generar herramientas que permitan detectar y contrarrestar situaciones de cyberbullying, para evitar que estas se presenten, o mitigar su impacto cuando se materialicen. Acciones que incluyen :

- Encuestas y estudios enfocados en indentificar el origen y el transfondo de esta problemátca tanto a nivel nacional como internacional, como el realizado por el *Instituto Nacional para la Evaluación de la Educación* (INEE), sobre disciplina, violencia y consumo de sustancias nocivas a la salud (García, Muñoz-Abundez & Martínez, 2007); la *Tercera Encuesta Nacional sobre Exclusión, Intolerancia y Violencia en las Escuelas de Educación Media Superior* (SEMS & SEP, 2014) realizada por la *Subsecretaría de Educación Media Superior (SEMS) y la Encuesta*

Nacional de Victimización del Crimen realizada por el *Departamento de Justicia de los Estados Unidos* y la *Oficina de Estadísticas de Justicia* (Lessne & Yanez, 2016) .

- Programas educativos antibullying orientados a la concientización social. Por ejemplo; KIVA, programa que se basa en el cambio de actitud y respuesta de los espectadores antes una situación de bullying (Williford et al., 2012); TEI "Tutoría entre iguales", mostrando TOLERANCIA CERO a actos de violencia o maltrato (Bellido, 2015); así como *Mybullying*, *NOHATE (No Hate Speech online)* y *ARBAX (Against racial bullying and xenophobia)* presentados en (Martínez, 2020).

- Así también, existen propuestas tecnológicas como MyFriends, Bully Freezone, Antibullying Sage, Antibullying-sector, Ray Chat y Treelp presentadas en (Universidad de Valencia, 2015); *MCDefender* desarrollado por los autores de (Vishwamitra et al., 2017) en 2017 y *At-Risk for Middle School Educators desarollado por los autores de* (Bradley & Kendall, 2019).

En el ámbito de la inteligencia artificial, y particularmente en el área de aprendizaje automático, múltiples estudios han aportado modelos basados en datos que permiten la identificación de mensajes con diferentes niveles de agresión o acoso. Sin embargo, en la revisión de la literatura, se observa que la mayoría de estas investigaciones se han enfocado en utilizar corpus de datos obtenidos de fuentes poco representativas, generalmente redes sociales abiertas como lo es Twitter. Una red social de este tipo permite obtener un volumen grande de datos de manera rápida y sencilla, pero presentan una deficiencia en su nivel de representatividad de escenarios reales de acoso cibernético; ya que utilizan conversaciones donde generalmente participan personas de diferentes rangos de edades y que utilizan palabras, frases o vocablos con diferentes significados derivado de provenir de diferentes zonas geográficas. Por lo tanto, aun cuando se reporte un buen desempeño al clasificar mensajes agresivos, los modelos generados no manejan una cobertura real del vocabulario utilizado por la audiencia objetivo.

Existe evidencia que en el entorno educativo se presenta en mayor medida situaciones de cyberbullying. Un número considerable de investigaciones han orientado su atención a atender a este problema que enfrentan jóvenes de los distintos niveles educativos. Sin embargo, el uso de Twitter como fuente de datos es la mayormente utilizada para generar modelos que detecten situaciones de acoso. En la presente investigación se sustenta que este tipo de red social es poco utilizada por la población objetivo, es decir jóvenes estudiantes, por lo que los conjuntos de datos carecen del lenguaje utilizado por los adolescentes, que está compuesto de palabras y frases utilizadas dentro de sus grupos de redes sociales privadas.

Adicionalmente, en la literatura se ha dado poca atención a reportar con detalle el proceso de creación del corpus, aportando escasa información sobre la fuente de datos, el mecanismo utilizado para su acopio, la fase de preprocesado y los criterios de etiquetado. En esta investigación se presenta el proceso de análisis y creación de un corpus con contenido de cyberbullying en idioma español, representativo de estudiantes mexicanos de nivel medio superior y superior, usando datos reales provenientes de grupos privados de las redes sociales

Facebook y WhatsApp. Se describe a detalle el origen de los datos, los instrumentos utilizados para su obtención, y las características consideradas para su organización y etiquetado; proceso que puede ser utilizado y replicado por otros investigadores para futuros estudios en el área.

El corpus generado toma en cuenta la fuente de origen de los datos y la representatividad de las conversaciones con presencia de cyberbullying, aun cuando el corpus final es menor en cantidad, comparado con corpus usados en otras investigaciones, la calidad de los datos es rica en comparación con aquellos corpus que se han obtenido con barrido web sin tener en cuenta la población objetivo y el ambiente en el que se han desarrollado. Con el objetivo de contrastar las características del corpus creado, en paridad con aquellos procedentes de redes libres, se presenta en esta investigación un comparativo del conjunto de características que influyen en la calidad (representatividad) de un corpus orientado a la detección de cyberbullying: como lo son la red social de origen, el tamaño del corpus, la elevada presencia de palabras únicas, el uso de lenguaje representativo de la audiencia destino, el proceso de etiquetado del contenido y un equilibrio en el número de muestras para cada una de las clases a clasificar.

El objetivo de la investigación se cumplió al utilizar una fuente de datos representativa en la generación de modelos para la detección de situaciones de cyberbullying; que puedan ser trasladados a entornos reales de operación, que permitirá una mejor precisión de los algoritmos a la hora de clasificar.

Apéndice A

Cuestionario

A.1 Cuestionario de preferencias de redes sociales

En este apéndice se presenta el cuestionario utilizado para recopilar información actual e identificar preferencias de redes sociales entre la población objetivo, estudiantes de los distintos niveles educativos.

CUESTIONARIO DE PREFERENCIAS EN REDES SOCIALES

Mi nombre es Karla Arce, egresada de la carrera de Ciencias Computacionales de UABC, actualmente me encuentro realizando mis estudios de Doctorado en el programa de Maestría y Doctorado en Ciencias e Ingeniería. El propósito del estudio de investigación es recopilar información actual que nos permita conocer las preferencias de redes sociales de los estudiantes de los distintos niveles educativos. El equipo de investigación que encabeza este estudio se especializa en temas referentes a computación y al desarrollo de tecnología de acuerdo al estudio de las necesidades de una comunidad. Su apoyo y retroalimentación es muy importante para nosotros. La presente encuesta tiene intereses estrictos de investigación y la información recopilada será de uso confidencial. Es importante mencionar que NO es objetivo de la encuesta evaluar a los participantes.

Grado académico:	Primaria ○	Secundaria ○	Preparatoria ○	Universidad ○	Posgrado ○
Escuela:	_____	**Turno:**	Matutino ○	Vespertino ○	Nocturno ○
Género:	Femenino ○	Masculino ○	**Edad:** _____		
A que te dedicas:	Estudio ○	Trabajo ○	Ambos ○		

¿Cuál es la red social que más utilizas?

Facebook	Twitter	Instagram	Snapchat	WhatsApp	Tumblr	Pinterest	Google+	Vine	Flickr	FourSquare	Otra:
○	○	○	○	○	○	○	○	○	○	○	_____

¿Por qué ? (Puedes seleccionar más de una opción)

Tengo más amigos en esta red social	Es mas fácil de usar	Es más segura	Rara vez tiene fallas	Es la más usada actualmente	Cuenta con muchas funcionalidades	Otra:
○	○	○	○	○	○	_____

Entre semana (Lunes-Viernes) ¿con qué frecuencia utilizas cada una de las siguientes redes sociales ?

REDES	Una hora al día	2-3 horas al día	3-6 horas al día	Más de 6hr al día	No la uso
Facebook	○	○	○	○	○
Twitter	○	○	○	○	○
Instagram	○	○	○	○	○
Snapchat	○	○	○	○	○
WhatsApp	○	○	○	○	○
Tumblr	○	○	○	○	○
Pinterest	○	○	○	○	○
Google+	○	○	○	○	○
Vine	○	○	○	○	○
Flickr	○	○	○	○	○
FourSquare	○	○	○	○	○
Otra:_____	○	○	○	○	○

Durante el fin de semana, ¿con qué frecuencia utilizas cada una de las siguientes redes sociales ?

REDES	Una hora al día	2-3 horas al día	3-6 horas al día	Más de 6hr al día	No la uso
Facebook	○	○	○	○	○
Twitter	○	○	○	○	○
Instagram	○	○	○	○	○
Snapchat	○	○	○	○	○
WhatsApp	○	○	○	○	○
Tumblr	○	○	○	○	○
Pinterest	○	○	○	○	○
Google+	○	○	○	○	○
Vine	○	○	○	○	○
Flickr	○	○	○	○	○
FourSquare	○	○	○	○	○
Otra:_____	○	○	○	○	○

Horario en el que utilizas en mayor medida tus redes sociales

	Mañana 6:00 am-11:59 am	Tarde 12:00 pm-5:59 pm	Noche 7:00 pm-11:59 pm	Muy noche 12:00 am-7:00am
Entre semana	○	○	○	○
Fin de semana	○	○	○	○

Dispositivo donde utilizas tus redes sociales con mayor frecuencia

Computadora ○ Celular ○ Tablet ○ Smart Watch ○

¿Qué actividades realizas con mayor frecuencia en tus redes sociales? (Coloca una cruz del lado izquierdo. Puedes seleccionar más de una opción). Ejemplo

X	Compartir imágenes

	Compartir imágenes
	Compartir videos
	Postear texto, comentarios o frases
	Chatear y enviar mensajes
	Ver videos
	Escuchar música
	Estar en contacto con mis amigos
	Jugar online

	Ver las actividades de mis contactos
	Publicar contenidos
	Seguir influencers
	Comentar contenido actual
	Fines profesionales
	Fines educativos
	Otra:

Apéndice B

Documento de etiquetado

B.1 Documento de etiquetado de conversaciones

En este apéndice se presenta el instrumento electrónico que se utilizó para el etiquetado de los diálogos, este documento permitió categorizar las conversaciones utilizando tres niveles de profundidad o granularidad: general, descriptiva y por enunciado.

ID	#Partipantes	Genero	Tópico	Tipo de agresión	Categoría bullying	Interacción	Nivel de bullying en la conversación (Escala 1-5)
Conv-1	2	2H/0M	Recreativo	Exclusión	Desempeño académico	1 a 1 (Uno a uno)	2- Agresión baja
0- Sin agresión	Usuario 2: mañana en donde sera entonces						
0- Sin agresión	Usuario 2: ?						
2- Agresión baja	Usuario 1: nadie le diga **(Detonador)**						
2- Agresión baja	Usuario 1: que se joda por no poner atencion						
1- Agresión muy baja	Usuario 2: jajajajaa						
0- Sin agresión	Usuario 2: casa del toño						
0- Sin agresión	Usuario 2: ya recorde jaja casa dle toño calimax						
1- Agresión muy baja	Usuario 1: ?? porque nadas preguntando entonces?						
2- Agresión baja	Usuario 1: usare mi dado e 20 caras para medir tu inteligencia						
0- Sin agresión	Usuario 1: 4						
Conv-2	2	2H/0M	Recreativo	Comentarios	Sexual	1 a 1 (Uno a uno)	2- Agresión baja
0- Sin agresión	Usuario 2: SAQUEN el manual						
0- Sin agresión	Usuario 2: xd						
0- Sin agresión	Usuario 1: wey						
2- Agresión baja	Usuario 1: quieres tampones?						
2- Agresión baja	Usuario 2: jajaaa noo						
2- Agresión baja	Usuario 1: te esta sngrando la chocha						
2- Agresión baja	Usuario 1: tapatela						
2- Agresión baja	Usuario 2: ajjajaja						
0- Sin agresión	Usuario 1: searcy lo tiene						
2- Agresión baja	Usuario 1: yo tambien pero no te lo voy a pasar						
0- Sin agresión	Usuario 2: jaja ok						
Conv-3	2	4H/0M	Recreativo	Comentarios	Sexual	1 a 1 (Uno a uno)	2- Agresión baja
0- Sin agresión	Usuario 1: esperate a la noce para que el toño diga si se arma en su cas						
0- Sin agresión	Usuario 2: ok						
0- Sin agresión	Usuario 2: vere jojos mietras xd						
0- Sin agresión	Usuario 3: Yo estoy con fullmetal alchemist y boku no hero						
2- Agresión baja	Usuario 1: yo estoy en tu hermana						
2- Agresión baja	Usuario 1: renate searcy best waifu						
2- Agresión baja	Usuario 4: Plox don't						
Conv-4	3	3H/0M	Recreativo	rtir información confi	Otro	1 a 1 (Uno a uno)	2- Agresión baja
0- Sin agresión	Usuario 1: sersi, tienes credito?						
0- Sin agresión	Usuario 2: Si						
0- Sin agresión	Usuario 3: marcale a @Usuario 4						
0- Sin agresión	Usuario 3: 6462377388						

Apéndice C

Diálogos obtenidos durante de la evaluación del sistema

C.1 Tabla con los diálogos realizados por los participantes durante la evaluación del sistema

En este apéndice se presenta una tabla con los diálogos realizados por los participantes durante la evaluación del sistema, así como la categorización asignada por el modelo a cada uno de los comentarios, la evaluación de la categorización por parte de los participantes en la conversación, y la evaluación de la categorización por parte del revisor adicional.

Comentario	Categorización	Participante 1	Participante 2	Participante 3	Revisor
En lo personal, apoyo este movimiento y creo que debería tener más visibilidad, por que aún en pleno 2021 se siguen viviendo muchas injusticias a parte de los comentarios que las mujeres tienen que "aguantar" por falta de educación de algunos hombres.	No-Bullying	Correcto	Autor	Correcto	Correcto
Yo estoy a favor pero no estoy de acuerdo en que las mujeres y hombres somos iguales	No-Bullying	Autor	Correcto	Correcto	Correcto
pienso que es una buena causa, sobretodo porque en paises como el nuestro la desigualdad de genero es mucho mas comun que en otros, y aun asi, hay personas que simplemente ignoran algunas situaciones que no son del todo igualitarias	No-Bullying	Correcto	Correcto	Autor	Correcto

no estoy de acuerdo, ellas ya tienen su lugar asignado en la sociedad	No-Bullying	Correcto	Incorrecto	Autor	Correcto
como que situaciones @jose7u7?	Bullying	Autor	Incorrecto	Incorrecto	Incorrecto
(comentemos cosas malas a ver si la pagina las marca como bullying)	No-Bullying	Correcto	Correcto	Autor	Correcto
creo que algunas feministas no saben como manifestarse y solo buscan hacer controversia, insultando y haciendo daños	No-Bullying	Autor	Correcto	Correcto	Correcto
concuerdo mi estimada, ademas de que justifican los daños causados con el argumento de que es una lucha contra los hombres y el patriarcado, siendo asi, donde se fue la igualdad?	No-Bullying	Correcto	Correcto	Autor	Correcto
el hecho de poner comentarios en contra del movimiento no significa que sean bullying, al final solo es una opinión (mientras no sean despectivos) xd	No-Bullying	Correcto	Autor	Correcto	Correcto
ademas de que en gran medida de la controversia del tema, es debido a la desinformacion, ya que hay gente que no tenemos muy claro que es en si el feminismo y las mismas que lo apoyan y confundimos terminos y las comparamos con el nazismo	Bullying	Incorrecto	Incorrecto	Autor	Correcto
pero se trata de dar tu opinión no	No-Bullying	Autor	Correcto	Correcto	Correcto

quedarte callada					
la mujer no sirve para nada, solo para criar hijos y cuidar la casa, que se vayan a cocinar mejor	Bullying	Correcto	Incorrecto	Correcto	Correcto
me parecen comentarios tontos, debemos de participar con sentido no haciendo comentarios nada mas por que si	Bullying	Autor	Incorrecto	Incorrecto	Correcto
solo estan para adornar la casa	No-Bullying	Correcto	Incorrecto	Autor	Incorrecto
(que paso ahi :'c)	No-Bullying	Correcto	Correcto	Autor	Correcto
el punto es hacer comentarios solo para saber si el sistema reconoce los que son despectivos	No-Bullying	Correcto	Autor	Correcto	Correcto
no leiste las instrucciones del correo o que	Bullying	Autor	Correcto	Incorrecto	Correcto
es el movimiento realmente malo, o la forma en la que es llevado a cabo es realmente lo malo?	Bullying	Incorrecto	Incorrecto	Autor	Incorrecto
lo tonto es afectar a otras personas por ejemplo cuando destruyen negocios buscando ser escuchadas	Bullying	Autor	Correcto	-	Correcto
y qué hay de las mujeres que son afectadas?	No-Bullying	Correcto	Autor	Correcto	Correcto
cómo propones que se den a escuchar entonces?	No-Bullying	Correcto	Autor	-	Correcto
claro que se debe de solucionar el problema, los feminicidios y las injusticias a las que las mujeres estan expuestas pero destruyendo no es la solucion	Bullying	Autor	Incorrecto	Incorrecto	Incorrecto
yo creo que se deberían de crear	No-Bullying	Autor	Incorrecto	Correcto	Correcto

alianzas y tratar de cambiar el concepto del feminismo, literalmente se están convirtiendo en feminazis					
por eso (no digo que sea la solución, solo es una forma de hacer presencia por que hablando no las escuchan), repito, entonces cuál es la solución para ti?	No-Bullying	Correcto	Autor	Correcto	Correcto
sabes que esas "alianzas" ya existen verdad? que el concepto del feminismo está claro al tratar de hacer un cambio de justicia, el problema es que existan personas como tú que clasifican un movimiento de justicia y amor en la misma etiqueta que los nazis, que SÍ fueron asesinos **si no me crees, busca páginas que se dedican a la educación sobre el feminismo en internet, no ocupas irte muy lejos**	No-Bullying	Incorrecto	Autor	Correcto	Correcto
y van a cambiar la justicia destruyendo el negocio de chuchita por el que ha trabajado muchos años?	No-Bullying	Autor	Correcto	Correcto	Correcto

Bibliografía

[1] Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access, 7, 70701-70718.

[2] Ansary, N. S. (2020). Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention. Aggression and violent behavior, 50, 101343.

[3] Aragón, M. A., Álvarez-Carmona, M., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., and Moctezuma, D. (2019). Overview of MEX-A3T at IberLEF 2019: Authorship and aggressive analysis in Mexican Spanish tweets. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), pp. 478-494.

[4] Aragón, M. E., Jarquín-Vásquez, H., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Gómez-Adorno, H., Posadas-Durán J. P., Bel-Enguix, G. (2020). Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressive analysis in Mexican Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), pp. 222-235

[5] Araujo, A., Pérez, J., & Rodriguez, W. (2018). Aplicación de una Red Neuronal Convolucional para el Reconocimiento de Personas a Través de la Voz. In Proc. Sexta Conferencia Nacional de Computación, Informática y Sistemas (pp. 77-81).

[6] Arce, J. (2014). Listado general de palabras en español. https://github.com/javierarce/palabras/find/master

[7] Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on twitter using Big Five and Dark Triad features. Personality and individual differences, 141, 252-257.

[8] Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. (2019, March). Detection of cyberbullying using deep neural network. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 604-607). IEEE.

[9] Bellido, A. G. (2015). Programa TEI "tutoría entre iguales". Innovación educativa, (25).

[10] Benavides, L. E. C. (2015). Una propuesta para identificar, clasificar y tipificar el Bullying (Acoso Escolar). Revista Iberoamericana para la Investigación y el Desarrollo Educativo ISSN: 2007-2619, (10).

[11] Berdugo-Gómez, N. (2020). Factores que influyen en la violencia escolar o bullying en adolescentes [Tesis de pregrado, Universidad Cooperativa de Colombia]. Repositorio Institucional UCC. https://repository.ucc.edu.co/handle/20.500.12494/18382

[12] Berlanga Silvente, V., & Vilà Baños, R. (2014). Cómo obtener un modelo de regresión logística binaria con SPSS. REIRE: revista d'innovació i recerca en educació.

[13] Bradley, E. G., & Kendall, B. (2019). Training teachers to identify and refer at-risk students through computer simulation. Journal of Technology in Behavioral Science, 4(4), 340-345.

[14] Bretschneider, U., Wöhner, T., & Peters, R. (2014). Detecting online harassment in social networks.

[15] Buss, A. (1961) The Psychology of Aggression. Editorial Wiley.

[16] Cerezo, F. (2009). Bullying: análisis de la situación en las aulas españolas. International Journal of Psychology and Psychological Therapy, 9(3), 383-394.

[17] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017a). Mean birds: Detecting aggression and bullying on twitter. In Proceedings of the 2017 ACM on web science conference, (pp. 13-22).

[18] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, b). Detecting aggressors and bullies on Twitter. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 767-768).
[19] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017, c). Measuring# gamergate: A tale of hate, sexism, and bullying. In Proceedings of the 26th international conference on world wide web companion (pp. 1285-1290).
[20] Chelmis, C., & Yao, M. (2019, June). Minority report: Cyberbullying prediction on Instagram. In Proceedings of the 10th ACM conference on web science (pp. 37-45).
[21] Chen, Y., Zhang, L., Michelony, A., & Zhang, Y. (2012). 4Is of social bully filtering: identity, inference, influence, and intervention. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2677-2679).
[22] Cheng, L., Guo, R., & Liu, H. (2019, May). Robust cyberbullying detection with causal interpretation. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 169-175).
[23] Cheng, L., Li, J., Silva, Y. N., Hall, D. L., & Liu, H. (2019, January). Xbully: Cyberbullying detection within a multi-modal context. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 339-347).
[24] Continente, X. G., Giménez, A. P., & Adell, M. N. (2010). Factores relacionados con el acoso escolar (bullying) en los adolescentes de Barcelona. Gaceta Sanitaria, 24, 103-108.
[25] Coob, J. G. C., & Marín, K. N. M. (2020). Psychometric properties and results of the school violence and bullying scale: how to distinguish bullying and school violence. Revista Electrónica de Psicología Iztacala, 23(3), 984-1014.
[26] Del Bosque, L. P., & Garza, S. E. (2016). Prediction of aggressive comments in social media: an exploratory study. IEEE Latin America Transactions, 14(7), 3474-3480.
[27] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), 1-30.
[28] El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.
[29] Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. IEEE Access 9: 103541-103563
[30] Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., ... & Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation, 55(3), 597-633.
[31] Farhadian, M., Shokouhi, P., & Torkzaban, P. (2020). A decision support system based on support vector machine for diagnosis of periodontal disease. BMC Research Notes, 13(1), 1-6.
[32] Fortunatus, M., Anthony, P., & Charters, S. (2020). Combining textual features to detect cyberbullying in social media posts. Procedia Computer Science, 176, 612-621.
[33] Gada, M., Damania, K., & Sankhe, S. (2021, January). Cyberbullying Detection using LSTM-CNN architecture and its applications. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.
[34] García, M. A., Muñoz Abundez, A., & Martínez, A. O. (2007). Disciplina, violencia y consumo de sustancias nocivas a la salud en las escuelas primarias y secundarias de México. Instituto Nacional para la Evaluación de la Educación.
[35] Garnacho, D. (2015). Dataset de Sentimientos en Español. https://github.com/garnachod/TwitterSentimentDataset
[36] Gegúndez Arias, M. E., & Pérez Borrero, I. (2021). Deep learning: fundamentos, teoría y aplicación. Deep learning, 1-261.

[37] Giumetti, G. W., & Kowalski, R. M. (2016). Cyberbullying matters: Examining the incremental impact of cyberbullying on outcomes over and above traditional bullying in North America. In Cyberbullying across the globe (pp. 117-130). Springer, Cham.
[38] Golze, J., Zourlidou, S., & Sester, M. (2020). Traffic Regulator Detection Using GPS Trajectories. KN-Journal of Cartography and Geographic Information, 70(3), 95-105.
[39] Herkama S, Salmivalli C. Kiva antibullying program. What is the KiVa antibullying program? Reducing cyberbullying in schools. International Evidence-Based Best Practices [Internet]. 2018 [acceso julio 2018]; 125-134. Disponible en: https://doi.org/10.1016/B978-0-12-811423-0.00009-2
[40] Hernández, C. P. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. Estudios de Lingüística del Español (ELiEs), (18), 1.
[41] Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.
[42] Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), 832-844.
[43] Instituto Nacional de Estadística y Geografía (INEGI). (2014). Encuesta de cohesión social para la prevención de la violencia y la delincuencia 2014.
[44] Juvonen, J., & Graham, S. (2014). Bullying in schools: The power of bullies and the plight of victims. Annual review of psychology, 65, 159-185.
[45] Karpathy, A., Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128-3137
[46] Kasture, A. S. (2015). A predictive model to detect online cyberbullying (Doctoral dissertation, Auckland University of Technology).
[47] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751.
[48] Kocatürk, M., & Türk-Kurtça, T. (2020). Moral Disengagement, Attitudes towards Violence and Irrational Beliefs as Predictors of Bullying Cognition in Adolescence. International Education Studies, 13(10), 47-59.
[49] Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). Detecting cyberbullying: query terms and techniques. In Proceedings of the 5th annual acm web science conference (pp. 195-204).
[50] Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications, 78(17), 24103-24119.
[51] Kurniasih, N., Kuswarno, E., Yanto, A., & Suganda, T. (2020). Science Mapping for Popular Topics in Cyberbullying Prevention Articles. Library Philosophy and Practice, 1-10.
[52] Larochelle, M. A., & Khoury, R. (2020, December). Generalisation of cyberbullying detection. In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 296-300). IEEE.
[53] Larrañaga, P., Inza, I., & Moujahid, A. (1997). Tema 6. Clasificadores Bayesianos. Departamento de Ciencias de la Computación e Inteligencia Artificial–Universidad del País Vasco-Euskal Herriko Unibertsitatea.
[54] Lessne, D., & Yanez, C. (2016). Student Reports of Bullying: Results from the 2015 School Crime Supplement to the National Crime Victimization Survey. Web Tables. NCES 2017-015. National Center for Education Statistics.
[55] Liu, J., & Graves, N. (2011). Childhood bullying: A review of constructs, concepts, and nursing implications. Public health nursing, 28(6), 556-568.

[56] Loredo-Abdalá, A., Perea-Martínez, A., & López-Navarrete, G. E. (2008). “Bullying”: acoso esocolar. La violencia entre iguales. Problemática real en adolescentes. Acta pediátrica de México, 29(4), 210-214.
[57] Loredo-Abdalá, A., Perea-Martínez, A., & López-Navarrete, G. E. (2008). “Bullying”: acoso esocolar. La violencia entre iguales. Problemática real en adolescentes. Acta pediátrica de México, 29(4), 210-214.
[58] Maimon, O. & Rokach, L. (2015). Data mining and knowledge discovery handbook.
[59] Margono, H., Yi, X., & Raikundalia, G. K. (2014). Mining Indonesian cyber bullying patterns in social networks. In Proceedings of the Thirty-Seventh Australasian Computer Science Conference-Volume 147 (pp. 115-124).
[60] Martínez, A. S. (2020). Efectividad de los programas de prevención de acoso escolar en las escuelas. NPunto, 3(27), 58-78.
[61] MathWorks. (s.f.) Support Vector Machine. Hiperplanos óptimos como límites de decisión. Recuperado el 19 de mayo del 2020 de https://es.mathworks.com/discovery/support-vector-machine.html
[62] Mayo, Matthew, (2017). “A General Approach to Preprocessing Text Data.” KDnuggets. Retrieved at Novemver 28, 2021, from the website https://www.k https://www.kdnuggets.com/2017/12/general-approach-preprocessing-text-data.html?fbclid=IwAR1McyxLzuPeKvz4iVUKq4W4RB9siBC202i7xgI7ERGDKdu5U62y8f8aEDMdnuggets.com/2017/12/general-approach-preprocessing-text-
[63] Miramontes, O. (1999). Los sistemas complejos como instrumentos de conocimiento y transformación del mundo. Perspectivas en las teorías de sistemas, 83-92.
[64] Modecki, K. L., Minchin, J., Harbaugh, A. G., Guerra, N. G., & Runions, K. C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. Journal of Adolescent Health, 55(5), 602-611.
[65] Mouheb, D., Abushamleh, M. H., Abushamleh, M. H., Al Aghbari, Z., & Kamel, I. (2019, June). Real-time detection of cyberbullying in arabic twitter streams. In 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS) (pp. 1-5). IEEE.
[66] Nigro, O., Xodo, D., Corti, G., & Terren, D. (2004). KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario. In VI Workshop de Investigadores en Ciencias de la Computación.
[67] Núñez-Prado, C. J., Chanona-Hernández, L., & Sidorov, G. (2020). Generación de un corpus en español con expresiones agresivas. Res. Comput. Sci., 149(8), 1055-1060.
[68] Oliveros, M., Figueroa, L., Mayorga, G., Cano, G., Quispe, Y., & Barrientos, A. (2009). Intimidación en colegios estatales de secundaria del Perú.
[69] Olweus, D. (1983). Low school achievement and aggressive behavior in adolescent boys. Human development: An interactional perspective, 353-365.
[70] Olweus, D. (1997). Bully/victim problems in school: Facts and intervention. European journal of psychology of education, 12(4), 495-510.
[71] Pacheco Leal, S. D., Díaz Ortíz, L. G., & García Flores, R. (2005). El clasificador Naïve Bayes en la extracción de conocimiento de bases de datos. Ingenierías, 7(27), 24-33.
[72] Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. RLA. Revista de lingüística teórica y aplicada, 46(1), 93-119.
[73] Pascanu, R., Mikolov, T. & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on Machine Learning, in PMLR 28(3):1310-1318.
[74] Perera, A., & Pumudu, F. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605-611.
[75] Pérez Carrasco, J. A., Serrano Gotarredona, M. D. C., Acha Piñero, B., Serrano Gotarredona, M. T., & Linares Barranco, B. (2011). Red neuronal convolucional rápida sin fotogramas

para reconocimientos de dígitos. In XXVI Simposio de la URSI (2011), p 1-4. Unión Científica Internacional de Radio.
[76] Pérez Hernández, M. C. (2002). Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. Estudios de lingüística del español, 18, 000-0.
[77] Pericàs, J. V., & Olive, J. M. (1999). Muestreo y recogida de datos en el análisis de redes sociales. Qüestiió: quaderns d'estadística i investigació operativa, 507-524.
[78] Potha, N., & Maragoudakis, M. (2014). Cyberbullying detection using time series modeling. In 2014 IEEE International Conference on Data Mining Workshop, 373-382. IEEE.
[79] Pyle, D. (1999). Data preparation for data mining. morgan kaufmann.
[80] Quiroz, G. (2007). Preparación y procesamiento de un corpus para la creación de materiales en la clase de español para propósitos específicos. In Actas del X Congreso Brasileño de Profesores de Español (pp. 131-150).
[81] Raisi, E., & Huang, B. (2016). Cyberbullying identification using participant-vocabulary consistency. arXiv preprint arXiv:1606.08084.
[82] Reisen, A., Viana, M. C., & dos Santos Neto, E. T. (2019). Adverse childhood experiences and bullying in late adolescence in a metropolitan region of Brazil. Child abuse & neglect, 92, 146-156.
[83] Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. Revista Ibérica de Sistemas e Tecnologias de Informação, (E28), 586-599.
[84] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.
[85] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.
[86] Ruiz-Ramírez, R., Pérez-Olvera, A., Zapata-Martelo, E., & Martínez-Corona, B. (2020). Análisis del bullying en tres escuelas del nivel medio superior. CPU-e, Revista de Investigación Educativa, (31), 28-50.
[87] Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
[88] Sak, H., Senior, A., Rao, K., and Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947
[89] Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., and Mao, M. (2014). Sequence discriminative distributed training of long shortterm memory recurrent neural networks. In Fifteenth annual conference of the international speech communication association.
[90] Salama, M., Kader, H. A., & Abdelwahab, A. (2021). An analytic framework for enhancing the performance of big heterogeneous data analysis. International Journal of Engineering Business Management, 13, 1847979021990523.
[91] Samghabadi, N. S., Monroy, A. P. L., & Solorio, T. (2020, May). Detecting early signs of cyberbullying in social media. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 144-149).
[92] Satapathy, S. C., Govardhan, A., Raju, K. S., & Mandal, J. K. (Eds.). (2015). Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Cham: Springer International Publishing. 2014
[93] Secretaría de Educación Pública (SEP), 2019. Programa Nacional de Convivencia Escolar. Coordinación Estatal del Programa Nacional de Convivencia Escolar.

https://www.gob.mx/escuelalibredeacoso/articulos/programa-nacional-de-convivencia-escolar-120992
[94] Simeone, O. (2018). A very brief introduction to machine learning with applications to communication systems. IEEE Transactions on Cognitive Communications and Networking, 4(4), 648-664.
[95] Subsecretaría de Educación Media Superior y Secretaría de Educación Pública. (2014). Tercera Encuesta Nacional sobre Exclusión, Intolerancia y Violencia en Escuelas de Educación Media Superior. Reporte Temático.
[96] Székely, M. (2008) "1a Encuesta Nacional Exclusión, Intolerancia y Violencia en Escuelas Públicas de Educación Media Superior", Secretaría de Educación Pública, México.
[97] Tapia, F., Aguinaga, C., & Luje, R. (2018, October). Detection of behavior patterns through social networks like twitter, using data mining techniques as a method to detect cyberbullying. In 2018 7th International Conference On Software Process Improvement (CIMPS) (pp. 111-118). IEEE.
[98] Torres, J. (2018). DEEP LEARNING Introducción práctica con Keras. Lulu. com.
[99] Universidad Internacional de Valencia. (2015). La tecnología como herramienta en la prevención del bullying. [Online]. Available: http://www.viu.es/la-tecnologia- como-herramienta- en-la-%0Aprevencion-del- bullying/.
[100] Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. Language Resources and Evaluation, 54(4), 851-874.
[101] Varela Barraza, J. A., Cabrera González, F., Zarabozo Enríquez de Rivera, D., Larios Villa, Y., & González Ortiz, M. (2013). Las 5000 palabras más frecuentes en los libros de texto oficiales de la educación básica en México. Revista electrónica de investigación educativa, 15(3), 114-120. Recuperado de http://redie.uabc.mx/vol15no3/contenido-varelaetal.html
[102] Vishwamitra, N., Zhang, X., Tong, J., Hu, H., Luo, F., Kowalski, R., & Mazer, J. (2017). MCDefender: Toward effective cyberbullying defense in mobile online social networks. In Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics, 37-42.
[103] Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020, July). Multi-modal cyberbullying detection on social networks. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
[104] Williford, A., Boulton, A., Noland, B., Little, T. D., Kärnä, A., & Salmivalli, C. (2012). Effects of the KiVa anti-bullying program on adolescents' depression, anxiety, and perception of peers. Journal of abnormal child psychology, 40(2), 289-300.
[105] Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012). Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, 656-666.
[106] Yao, M., Chelmis, C., & Zois, D. S. (2019, May). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In The World Wide Web Conference (pp. 3427-3433).
[107] Zillmann, D. (1979). Hostility and Aggression. Editorial Taylor and Francis.