Automatic Cyberbullying Detection: A Mexican Case in High School and Higher Education Students

K. I. Arce-Ruelas, O. Álvarez-Xochihua, L. Pelegrin, L. Cardoza-Avendaño and J. A. González-Fraga

***Abstract*—The social interaction among young students has been partially or totally transformed to mobile-based communication, specifically through the use of social networks. This new communication environment has allowed a more immediate, diverse and massive interaction, offering a faster and more effective situation when carrying out academic and recreational activities. However, this scenario has also promoted the phenomenon of social harassment known as bullying, exponentially increasing its scope and diversifying the types and forms of aggression. Machine learning and natural language processing techniques have been used to create models that detect bullying situations among students, using data corpus from mainly public social networks. However, generally, these data sources are not representative of the social networks commonly used by the students; generating classification models that do not consider the vocabulary used by this social group. This article describes the methodology used to create a representative data corpus of the interaction between Mexican high school and university students, and a comparative analysis on characteristics that influence the quality of the content of a corpus in this domain. In addition, the performance achieved by implementing various machine learning models to identify bullying situations is presented. The best result is reported for the Naive Bayesian classifier (F1-Score of 0.862), performing better than models based on deep learning such as Recurrent (F1-Score of 0.845) and Convolutional (F1-Score of 0.807) Neural Networks.**

***Index Terms*— Bullying, cyberbullying, machine learning, social networks, deep learning**

I. Introducción

A lo largo de la historia humana, 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 [1].

K.I. Arce-Ruelas, Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California, B.C., México. karla.arce@uabc.edu.mx

O. Álvarez-Xochihua, Facultad de Ciencias, Universidad Autónoma de Baja California, B.C., México. aomar@uabc.edu.mx (corresponding author)

L. Pellegrin, Facultad de Ciencias, Universidad Autónoma de Baja California, B.C., México. luis.pellegrin@uabc.edu.mx

L. Cardoza-Avendaño, Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California, B. C., México. lcardoza@uabc.edu.mx

J. A. González-Fraga, Facultad de Ciencias, Universidad Autónoma de Baja California, B. C., México. angel_fraga@uabc.edu.mx.

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, las interacciones agresivas se categorizan como un tipo de acoso psicológico o moral, al cual se le refiere internacionalmente mediante el término en inglés *bullying* (acoso). En la literatura existen diversas definiciones del concepto de *bullying*. Existen elementos homogeneizados entre estas definiciones que hacen referencia a: 1) 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 [2].

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 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, el acoso escolar ya no se limita a la agresión presencial realizada durante las horas de escuela, sino que se ha materializado 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 [3]; 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* o ciberacoso en español.

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 [4-5], se pone en evidencia la importancia de atender este problema a través de la misma tecnología donde se genera. Se deben crear y adaptar desarrollos tecnológicos que permitan detectar, clasificar, predecir y, de ser posible, evitar escenarios de cyberbullying.

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 [6]. Mediante técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés), y algoritmos propios de aprendizaje automático, 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 cyberbullying con diferentes niveles de precisión, que van del rango de 0.45 hasta un 0.95 [7], la variación en los resultados deriva principalmente por el algoritmo utilizado y el origen y las características del corpus de datos. Muchos de estos estudios están orientados a detectar o predecir escenarios de ciberacoso entre jóvenes en entornos educativos de nivel medio superior y superior. Sin embargo, la mayoría de las investigaciones utilizan corpus lingüísticos con contenido agresivo provenientes de fuentes de acceso libre, 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) y de conversaciones acompañadas de datos multimodales (audio, imágenes, memes y/o videos).

En el presente artículo se describe el proceso de creación y análisis de un corpus de *cyberbullying* en idioma español mexicano, donde se toman 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. Considerando la desventaja de contar con un corpus de tamaño reducido, pero con calidad en su contenido, se reporta un desempeño aceptable comparado con lo reportado en la literatura.

Primero, se describe el resultado de la revisión de trabajos relacionados al problema de estudio. Posteriormente, se presenta la metodología utilizada para la creación de un corpus de datos multimodal en el idioma español mexicano con presencia de *cyberbullying*. Se continúa con la presentación de los modelos de clasificación implementados utilizando algoritmos tradicionales de aprendizaje automático y de aprendizaje profundo. Finalmente se deriva un análisis de los resultados y las conclusiones del estudio realizado.

II. Trabajos relacionados

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 [7]. La explotación de redes sociales, como *Twitter* [8], ha detonado la generación de grandes bancos de datos para este fin. Estas fuentes de información son comúnmente utilizadas para generar modelos que permiten detectar o predecir 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 [9]), que incluyen interacciones entre personas de diversos círculos sociales y edades. Se construyen corpus usando diccionarios de palabras vulgares y no coloquiales [10-12] para filtrar *tweets* con interacciones de interés. Desafortunadamente, el contexto de interacción de estas fuentes no es el de 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 [13].

En [7] se presenta un análisis detallado de 22 estudios 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 I se presentan las principales características evaluadas en el estudio.

TABLA I
CARACTERÍSTICAS DE 22 CORPUS UTILIZADOS EN LA GENERACIÓN DE MODELOS 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

En el estudio se identifican como las principales fuentes de información las redes sociales *Twitter* y *YouTube*, obteniendo los datos mediante 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 utilizan la red social *Twitter* como fuente principal de datos y todas ellas consideran 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 [14-26] (ver Tabla II).

TABLA II
CARACTERÍSTICAS DE 13 CORPUS UTILIZADOS EN LA GENERACIÓN DE MODELOS 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), un extracto de estos resultados considerando las métricas más utilizadas se presenta en la Tabla III.

TABLA III
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 la 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; destacando un mejor desempeño al utilizar RF y CNN.

En otra revisión de literatura, presentada recientemente en [44], se ratifica la baja atención en las características mencionadas al crear corpus para el análisis de cyberbullying, y se menciona como un problema adicional el desbalance entre las muestras positivas (con *bullying*) y negativas (sin *bullying*). 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*. El corpus propuesto en la presente investigación cuenta con un 60% de muestras positivas, y un 40% de muestra negativas por lo que se considera como un corpus con un nivel de balance adecuado (ver Tabla IV).

TABLA IV
BALANCE DE MUESTRAS POSITIVAS (CON-BULLYING) Y NEGATIVAS (SIN-BULLYING) EN CORPUS

Corpus	Positivas (Con Bullying)	Negativas (Sin-Bullying)
Propio	60%	40%
30%	10%	90%
36%	11%-29%	89%-71%
30%	30%-39%	70%-61%
4%	42%	58%

* Porcentajes con base en 24 corpus analizados en [44] y el propio.

III. 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: adolescentes estudiantes mexicanos de nivel medio superior y superior. Considerando conversaciones donde interactúan distintos números 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 es 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 se siguió una metodología con base en el proceso de *Descubrimiento de Conocimiento en Bases de Datos* (KDD, por sus siglas en inglés) [27]. 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.*

1) *Selección 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 [1]; particularmente en los niveles medio superior y superior [28-30]. 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 [27]. En el presente trabajo se procedió a identificar una fuente de datos representativa de los

sujetos de estudio seleccionados, y posteriormente a recopilar los datos de interés, proceso descrito 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 V, siendo evidente que la representatividad de *Twitter*, red social comúnmente utilizada en otros estudios, es mínima o casi nula en estos niveles educativos.

TABLA V
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.

2) *Acopio de conversaciones.*

Se procedió a la creación del corpus objetivo invitando a participar a estudiantes de tres grupos 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 la misma aplicación de red social, o 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 inicial. 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 VI.

TABLA VI
CONCENTRADO DE DIÁLOGOS CON PRESENCIA DE *CYBERBULLYING*

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

3) *Preprocesamiento: 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 procesamiento. Para esto se tomaron en cuenta las siguientes consideraciones.

En la literatura se indica que una oración usada para ofender o agredir consiste en 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 [31]. 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 [31]:

a) Peyorativas: aquellas palabras o frases que indican una idea desfavorable o despectiva.
b) Obscenas: palabras o frases ofensivas al pudor, generalmente de contenido sexual.
c) 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: eliminando diálogos incompletos, descartando conversaciones que no presentaran un escenario de una interacción real, o 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) *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 adecuada de los datos (ej. uso de categorías en lugar de valores continuos) y reducción de dimensiones (ej. eliminando o reduciendo datos o registros incompletos) [27]. Lo anterior permite obtener una representación de las características útiles para el proceso de minado.

En esta etapa del estudio, 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 Fig. 1 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…

Fig. 1. 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*:

a) *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) *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 [32-36], con el aval de 3 profesionales en el área de psicología educativa.
c) *Por enunciado*, finalmente, cada enunciado en las conversaciones fue clasificado con base en su nivel de agresión, utilizando la 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. Posterior al proceso de etiquetado, se generó un solo documento que concentró el clasificado 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 los etiquetadores 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.

IV. Características del Corpus

En esta sección se describen las características del contenido del corpus lingüístico y se realiza un comparativo con dos corpus en español obtenidos de la red social Twitter [37], fuente comúnmente utilizada en estudios previos.

A. Características del Corpus

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. Algunas observaciones identificadas 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 1).

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*). 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%) y de muchos a 1 (19%) (Fig. 2).

Aun cuando se observaron palabras y frases fuertes o altisonantes, al 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.

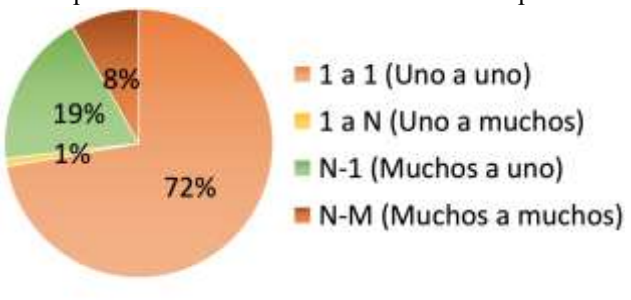

Fig. 2. Interacción entre participantes de conversaciones.

B. Análisis Comparativo entre 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 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) [37]. En este análisis, primeramente, se comparó el comportamiento de frecuencias entre palabras. Ya que la existencia de palabras poco frecuentes tiende 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 [38], 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 el uso de palabras frecuentes. Para este último aspecto se utilizó un conjunto de datos consistente en las 1,000 palabras más utilizadas en el idioma español mexicano, elaborado por [39]. En la Tabla VII se presentan los resultados obtenidos.

TABLA VII
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 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 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 los autores de la presente investigación, destacamos la presencia de tres características fundamentales, enfatizadas en la revisión de literatura sobre *cyberbullying* realizada por [7, 44]: red social de procedencia, balance del corpus y el etiquetado por expertos.

V. 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.

A. Algoritmos de Aprendizaje Automático Utilizados

Inicialmente, se consideraron los siguientes 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 una variante para modelar espacios no lineales con 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 [40]. 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 [41]. 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.

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 [40]. Esta modalidad de red neuronal también ha sido experimentada con éxito en el dominio de NLP [42]. 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 se generó ejecutando 5 iteraciones de entrenamiento con una manipulación por lotes de 32 sentencias. Compilando con la función de pérdida de entropía cruzada binaria y el optimizador *Adam* (Estimación Adaptativa de Momentos).

B. *Preprocesado de los Datos*

En investigaciones como la presentada en este artículo 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 por lo 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 [43].

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*, y 6) obtención de la raíz de las palabras (*stemming*).

C. *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 calidad para construir modelos de clasificación para identificar expresiones de *cyberbullying* con un nivel de confianza aceptable. Se procedió a generar los modelos de clasificación usando los principales algoritmos de aprendizaje automático reportados en la literatura. Se 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. 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 VIII se presenta el desempeño obtenido por cada uno de los algoritmos utilizados. Considerando como referencia 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 VIII
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 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 IX).

TABLA IX
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	
[14]	V-F			**0.929**				
[7]	V-F	0.740	0.750	0.650				
[15]	EXAC						**0.939**	
[17]	V-F	**0.868**						
[18]	EXAC			0.740	0.644			
[23]	V-F		0.760					
[24]	V-F							0.860
[25]	V-F							**0.866**

* PROP (método de clasificación propuesto por los autores)

Se observa que 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 [7 y 18] y ligeramente inferior a la reportada en [14]. Es importante mencionar que el menor desempeño obtenido en [7] 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 también 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, aun 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 [15]. 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 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 hombres participando en esa conversación, y
5) *Participantes_masculinos*, valor numérico que indica la cantidad de mujeres participando en esa conversación.

Mediante una prueba *Chi-cuadrada*, usando la clase *SelectKBest* de *scikit-learn*, las siguientes tres características fueron seleccionadas para el experimento, al resultar con un mejor nivel de predicción 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 X). Se considera como trabajo futuro realizar un estudio más detallado sobre el análisis multivariable para la identificación de cyberbullying, considerando lo estipulado en [44].

TABLA X
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

VI. CONCLUSIONES

Los resultados obtenidos en esta investigación permitieron generar un corpus de datos con contenido de *cyberbullying* representativo de estudiantes de habla hispana mexicana de nivel medio superior y superior. Se describe a detalle el origen de los datos, los instrumentos utilizados para su obtención, y las características consideradas para su organización y etiquetado.

Se identificaron un 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.

Los experimentos realizados permiten observar 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 incidieron 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 en la generación de modelos como características que permitan extender el análisis multivariable presentado en la presente investigación.

REFERENCES

[1] A. Loredo-Abdalá, A. Perea-Martínez, & G. López-Navarrete, "'Bullying': acoso escolar. La violencia entre iguales. Problemática real en adolescentes", Acta Pediátrica de México, vol. 29, no. 4, pp. 210–4, 2008.

[2] X. Garcia Continente, A. Pérez Giménez, & M. Nebot Adell, "Factores relacionados con el acoso escolar (bullying) en los adolescentes de Barcelona", Gaceta Sanitaria, vol. 24, no. 2, pp. 103–108, 2010.

[3] L. E. C. Benavides, "Una propuesta para identificar, clasificar y tipificar el Bullying (Acoso Escolar)". *Revista Iberoamericana para la Investigación y el Desarrollo Educativo* ISSN:2007-2619, no. 10, 2015.

[4] D. Lessne & C. Yanez, "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*, 2016.

[5] K. L. Modecki, J. Minchin, A. G. Harbaugh, N. G. Guerra & K. C. Runions, "Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying". *Journal of Adolescent Health*, vol. 55, no. 5, pp. 602-611, 2014.

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

[7] H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J. P. Carvalho, S. Oliveira & I. Trancoso, "Automatic cyberbullying detection: *A systematic review". Computers in Human Behavior*, vol. 93, pp. 333-345, 2019.

[8] K. Dinakar, B. Jones, C. Havasi, H. Lieberman & R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying". *In: ACM Transactions on Interactive Intelligent Systems* (TiiS), vol. 2, no.. 3, pp. 18, 2012.

[9] J. M. Xu, K. S. Jun, X. Zhu & A. Bellmore, "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*. NAACL HLT '12, pp. 656–666, 2012.

[10] C. Núñez-Prado, L. Chanona-Hernández & G. Sidorov, "Generation of a Corpus in Spanish with Aggressive Expressions". *Research in Computing Science* 149(8), pp. 1055-1060, 2020.

[11] M. A. Aragón, M. Álvarez-Carmona, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda & D. Moctezuma, "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, 2019.

[12] M. E. Aragón, H. Jarquín-Vásquez, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, J. P. Posadas-Durán & G. Bel-Enguix, "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, 2020.

[13] F. Cerezo, "Bullying: análisis de la situación en las aulas españolas". International Journal of Psychology and Psychological Therapy, vol. 9, no. 3, pp. 383-394, 2009.

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

[15] V. Banerjee, J. Telavane, P. Gaikwad & P. Vartak, "Detection of Cyberbullying Using Deep Neural Network". *In 2019 5th International Conference on Advanced Computing & Communication Systems* (ICACCS), *IEEE*, pp. 604-607, 2019.

[16] F. Tapia, C. Aguinaga & R. Luje, "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), IEEE*, pp. 111-118, 2018.

[17] C. Chelmis & M. Yao, "Minority Report: Cyberbullying Prediction on Instagram". *In Proceedings of the 10th ACM Conference on Web Science*, pp. 37-45, 2019.

[18] A. Kumar & G. Garg, "Sentiment analysis of multimodal twitter data". Multimedia Tools and Applications, vol.78, no.17, pp.24103-24119, 2019.

[19] D. Mouheb, M. H. Abushamleh, Z. Al Aghbari & I. Kamel, "Real-time detection of cyberbullying in arabic twitter streams". *In 2019 10th IFIP International Conference on New Technologies, Mobility and Security* (NTMS), *IEEE*, pp. 1-5, 2019.

[20] L. Cheng, R. Guo & H. Liu, "Robust cyberbullying detection with causal interpretation". In Companion Proceedings of the 2019 World Wide Web Conference, pp. 169-175, 2019.

[21] L. Cheng, J. Li, Y. N. Silva, D. L. Hall & H. Liu, "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, 2019.

[22] M. Yao, C. Chelmis & D. S. Zois, "Cyberbullying 0065nds here: Towards robust detection of cyberbullying in social media". *In: The World Wide Web Conference*, pp. 3427-3433, 2019.

[23] N. S. Samghabadi, A. P. L. Monroy & T. Solorio, "Detecting Early Signs of Cyberbullying in Social Media". In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 144-149, 2020.

[24] K. Wang, Q. Xiong, C. Wu, M. Gao & Y. Yu, "Multi-modal cyberbullying detection on social networks". *In 2020 International Joint Conference on Neural Networks* (IJCNN), *IEEE*, pp. 1-8, 2020.

[25] M. Fortunatus, P. Anthony & S. Charters, "Combining textual features to detect cyberbullying in social media posts". *Procedia Computer Science*, vol. 176, pp. 612-621, 2020.

[26] 24 D. Van Bruwaene, Q. Huang & D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media". *Language Resources and Evaluation*, pp. 1-24, 2020.

[27] O. Maimon & L. Rokach, "Data mining and knowledge discovery handbook", 2015.

[28] Instituto Nacional de Estadística y Geografía (INEGI). "Encuesta de cohesión social para la prevención de la violencia y la delincuencia 2014", 2014.

[29] M. Kocatürk & T. Türk-Kurtça, Moral Disengagement, "Attitudes Towards Violence and Irrational Beliefs as Predictors of Bullying Cognition in Adolescence". s*International Education Studie*, vol. 13, no. 10, 2020.

[30] A. Reisen, M. C. Viana & E. T. dos Santos Neto, "Adverse childhood experiences and bullying in late adolescence in a metropolitan region of Brazil". Child abuse & neglect, vol. 92, pp. 146-156, 2019.

[31] S.C. Satapathy, A. Govardhan, K.S. Raju & J. K. Mandal, "Emerging ICT for Bridging the Future". Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) 1(Vol.337), Springer, 2014.

[32] N. Kurniasih, E. Kuswarno, A. Yanto & T. Suganda, "Science Mapping for Popular Topics in Cyberbullying Prevention Articles". *Library Philosophy and Practice* (e-journal), pp. 1-10, 2020.

[33] N. S. Ansary, "Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention*". Aggression and violent behavior*, vol. 50, pp. 101343, 2020.

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

[35] R. Ruiz-Ramírez, A. Pérez-Olvera, E. Zapata-Martelo & B. Martínez-Corona," Análisis del bullying en tres escuelas del nivel medio superior". CPU-e, *Revista de Investigación Educativa*, vol. 31, pp. 28-50, 2020.

[36] K. N. M. Marín & J. G. C. Coob, "Psychometric properties and results of the school violence and bullying scale: how to distinguish bullying and school violence". *Revista Electrónica de Psicología Iztacala*, vol. 23, no.. 3, pp. 984-1014, 2020.

[37] D. Garnacho. "Dataset de Sentimientos en Español". https://github.com/garnachod/TwitterSentimentDataset

[38] J. Arce. "Listado general de palabras en español". https://github.com/javierarce/palabras/find/master

[39] J. A. Varela, F. Cabrera, D. Zarabozo, Y. Larios & M. González, "Las 5000 palabras más frecuentes en los libros de texto oficiales de la educación básica en México". Revista Electrónica de Investigación Educativa, vol.15, no.3, pp.114-123, 2013. Recuperado de http://redie.uabc.mx/vol15no3/contenido-varelaetal.html

[40] A. Karpathy, L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 3128-3137, 2015.

[41] R. Pascanu, T. Mikolov & Y. Bengio, "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning,* in PMLR vol. 28, no.3, pp. 1310-1318, 2013.

[42] Y. Kim, "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 1746–1751, 2014.

[43] M. Salama, H. A. Kader & A. Abdelwahab, "An analytic framework for enhancing the performance of big heterogeneous data analysis". *International Journal of Engineering Business Management*, vol. 13, pp. 1847979021990523, 2021.

[44] F. Elsafoury, S. Katsigiannis, Z. Pervez & N. Ramzan, "When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection". *IEEE Access* 9: pp. 103541-103563, 2021.

Karla Ivette Arce Ruelas es estudiante de Doctorado en el programa de Ciencias e Ingeniería de la Universidad Autónoma de Baja California, recibió los títulos de Licenciada y Maestra en Ciencias Computacionales, por la misma institución. Su interés de investigación se centra en el uso de tecnología en apoyo a la educación infantil, el modelado del conocimiento, Procesamiento del Lenguaje Natural y Aprendizaje Máquina.

Omar Álvarez-Xochihua es profesor investigador en la Universidad Autónoma de Baja California. En 2011, obtuvo el grado de Doctor en Ciencias Computacionales en Texas A&M University, USA. Su interés en investigación es en el área de tecnología educativa y desarrollo de Sistemas de Tutoría Inteligente y Procesamiento de Lenguaje Natural.

Luis Pellegrin es profesor-investigador en la Universidad Autónoma de Baja California. Recibió su título de Doctorado en Ciencias Computacionales por el Instituto Nacional de Astrofísica, Óptica y Electrónica en 2017. Sus áreas de interés están centradas en la interacción de Visión por Computadora y Procesamiento del Lenguaje Natural.

Liliana Cardoza Avendaño. Ingeniero en eléctrica, Maestría en Ingeniería con perfil en electrónica, Doctorado en ciencias. Miembro del Sistema Nacional de Investigadores nivel I desde 2014. Pertenece al Cuerpo Académico "Sistemas complejos y sus aplicaciones". Responsable del proyecto "sincronización de redes complejas. Línea de investigación: Comunicaciones seguras, sincronización de sistemas caóticos.

José Ángel González Fraga es profesor investigador en la Universidad Autónoma de Baja California, recibió el título de maestría y doctorado en Ciencias de la Computación por el Centro de Investigación Científica y de Educación Superior de Ensenada, México. Sus intereses de investigación incluyen, el Reconocimiento adaptativo de patrones, Visión por computadora, Tecnologías educativas y Robótica.