UNIVERSIDAD CUAUHTÉMOC
VANGUARDIA Y EXCELENCIA

EDUCACIÓN A DISTANCIA
UNIVERSIDAD CUAUHTÉMOC

ACUERDO NO. 1857 CON FECHA DEL 09 DE SEPTIEMBRE DE 2015 DEL INSTITUTO DE EDUCACIÓN DEL ESTADO DE AGUASCALIENTES

"ANÁLISIS CON TÉCNICAS DE BIG DATA ANALYTICS PARA MITIGAR LOS RIESGOS DEL CIBERBULLYNG EN LA ERA DIGITAL"

TESIS PARA: **MAESTRÍA EN CIENCIA DE LOS DATOS Y PROCESAMIENTO DE DATOS MASIVOS BIG-DATA**

PRESENTA(N): **LUIS ALBERTO OCAMPO LOMBANA**

DIRECTOR(A) DE TESIS: **DR. IVÁN CASTILLO ZUÑIGA**

Aguascalientes, julio 2019.

ASUNTO: Carta de autorización.

Aguascalientes, Ags., 11 de julio de 2019.

LIC. ROGELIO MARTÍNEZ BRIONES
UNIVERSIDAD CUAUHTÉMOC PLANTEL AGUASCALIENTES
RECTOR GENERAL

P R E S E N T E

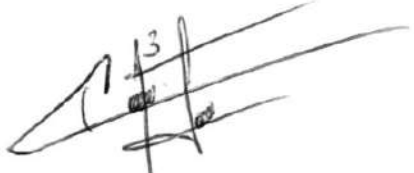
Por medio de la presente, me permito informar a Usted que he asesorado y revisado el trabajo de tesis titulado:

" ANÁLISIS CON TÉCNICAS DE BIG DATA ANALYTICS PARA MITIGAR LOS RIESGOS DEL CIBERBULLYING EN LA ERA DIGITAL "

Elaborado por el Ingeniero en Sistemas Computacionales **LUIS ALBERTO OCAMPO LOMBANA**, considerando que cubre los requisitos para poder ser presentado como trabajo recepcional para obtener el grado de Maestro en Ciencias de los Datos y Procesamiento de Datos Masivos (BIG-DATA).

Agradeciendo de antemano la atención que se sirva a dar la presente, quedo a sus apreciables órdenes.

ATENTAMENTE



Dr. Iván Castillo Zúñiga
Director de tesis

Plantel Aguascalientes

Acuerdo No. 1857 del 8 de abril del 2015 del
Instituto de Educación del Estado de Aguascalientes

Tesis.

Para Obtener el Grado de Maestría
en Ciencias de los Datos
y Procesamiento de Datos Masivos (BIG-DATA)

Título de la Tesis.

Análisis con Técnicas de Big Data Analytics para Mitigar los Riesgos del Ciberbullying en la Era Digital.

Presenta:

Luis Alberto Ocampo Lombana.

Director:

Dr. Iván Castillo Zúñiga.

CDMX, México 2019.

Índice

Índice de figuras.

Índice de tablas.

Glosario de términos.

Término	Descripción
Algoritmo	En Big Data, ayuda en la búsqueda de patrones y relaciones entre variables entre tanta cantidad de datos.
Algoritmo C.4.5	Es usado para generar arboles de decisión desde un grupo de datos de entrenamiento, el cual puede ser usado para clasificación y es referido como un clasificador estadístico.
Amenazas físicas	Enviar mensajes que impliquen amenazas a la seguridad física de una persona.
Análisis de Correlación	El Análisis de correlación es la herramienta estadística utilizada para estudiar la cercanía de la relación entre dos o más variables. Se dice que las variables están correlacionadas cuando el movimiento de una variable va acompañado por el movimiento de otra variable.
Análisis Descriptivo	Se trata del tipo de analítica más simple. El que permite condensar el Big Data en datos más pequeños, con piezas de información más manejable.
Analista de Datos	Es la persona dedicada profesionalmente a analizar, con técnicas estadísticas, los datos de la empresa para la que trabaja.

Analítica de negocio	Es la forma en la que un negocio utiliza sus técnicas para obtener información a través de sus datos. Se suele hacer mediante análisis estadísticos.
Analítica predictiva	Es una ciencia que se utiliza para predecir lo que puede pasar en el negocio, con base a esos datos históricos que utiliza técnicas como la del modelado predictivo basadas en algoritmos estadísticos y de aprendizaje automático.
Analytics	Es la forma de capturar, procesar y analizar información para que se conviertan en insights.
API (Application Programming Interface)	Una interfaz o “intermediario” que permite que dos programas de software interactúen entre sí. En el contexto de las operaciones de cambio, una API se refiere a la interfaz o plataforma que permite que su plataforma se conecte con el mercado.
BI (Business Inteligence)	Es el método que transforma informaciones almacenadas y analizadas en datos que son estratégicos para una empresa y que se convierten en ganancia para el negocio.
Blaze	Es una interfaz para consultar datos en diferentes sistemas de almacenamiento. Se le conoce también como una librería de Python (lenguaje de programación).
Bloquear	Acción de restringir la participación activa o el acceso de un usuario determinado en el entorno de Internet, y dentro de

	sistemas específicos como chats, redes sociales, foros… Ejemplo: “La víctima debe bloquear a los acosadores cuanto antes y no permitirles el acceso a sus datos privados”
Bofetadas de felicidad.	Grabar a alguien que está siendo acosado o intimidado de una manera que generalmente implica abuso físico, y luego publicar el video en línea para que sea visto por el público.
Bosque aleatorio	Es un método de ensamblado que utiliza varios algoritmos de aprendizaje.
BSON	Abreviatura de Binary JSON. Es un formato de datos de computadora utilizado principalmente como formato de almacenamiento de datos y tranferencia en la red, en la base de datos MongoDB.
Bully	El niño acosador es conocido como “niño bully”. Es el que comienza el hostigamiento y lo mantiene activamente gracias a sus secuaces y a aquellos que, sin ser activos, permiten o aprueban el acoso. Ejemplo: “El niño bully no empatiza con el dolor ajeno y tiende a la impulsividad”.
Bullying	Es el término inglés que se produce dentro del entorno del menor y entre niños de la misma edad. Su principal escenario es la escuela y deja secuelas a la víctima por ser un maltrato físico y psicológico. Ejemplo: “El bullying afecta a un gran número de adolescentes y tiene riesgo de suicidio”.

Chat	Lugar en internet donde se mantienen conversaciones entre dos o más usuarios de forma simultánea y en tiempo real. A veces son temáticos y otras son “macrochats” de alcance ilimitado como twitter. Por ello son una vía frecuente de acoso contra la víctima del Ciberbullying. Ejemplo: “El chat de la clase de literatura permite compartir los textos y consultar dudas a compañeros conocidos, pero es aburrido, mientras que el de las revistas de moda es más divertido aunque no conozcas a nadie”.
Ciberacoso o Ciberbullying.	El bullying llevado a término en el contexto de las nuevas tecnologías, vía internet, de forma que la víctima se ve acosada, humillada y avergonzada en cualquiera de sus dispositivos con acceso a Internet. Ejemplo: “El ciberacoso o Ciberbullying multiplica el efecto negativo en el niño”.
Cluster	Conjunto de servidores (o nodos) que permiten garantizar la continuidad del servicio y distribuir la carga de procesamiento/red.
Crawler o rastreador	Es un programa que analiza los documentos de un sitio Web creando una base de datos con la información encontrada.
Cyberdating.	Se trata de “quedar” o tener una cita, pero en lugar de en una cafetería, en un espacio virtual, como pueda ser un chat. Es decir, haciéndose uso de las TICs y en casi todos los casos llevándose a cabo entre jóvenes que se han conocido en

	Internet y no en persona. “El Cyberdating es una práctica peligrosa porque intimas con desconocidos que te pueden querer engañar”.
Data Mining	Es el proceso previo al Analytics, es la minería de los datos, la forma de descubrir informaciones relevantes.
Data Science (Ciencia de datos)	La oportunidad que los datos ofrecen para generar nuevo conocimiento requiere de técnicas sofisticadas de preparación de estos datos (estructuración) y análisis de los mismos. Así en Internet, sistemas de recomendación, traducción automática y otros sistemas de Inteligencia Artificial se basan en técnicas de Data Science.
Data Scientist	Es el analista de datos. La persona que capturará los insights, las principales informaciones dentro de un gran volumen de información.
Datos estructurados y no estructurados	Los estructurados tienen una organización lógica, pero con posibilidades pequeñas de extracción de informaciones para el Big Data. Por otro lado, los no estructurados son desorganizados, como los mensajes en emails y redes sociales, pero permiten una gran posibilidad de extracción de insights comerciales.
Datos sensibles	Se refiere a la información personal del usuario, relativa a diferentes cuestiones que abarcan desde el origen racial hasta

	las convicciones políticas, religiosas o morales, además de aspectos referidos a las preferencias sexuales, al estilo de vida o a la salud. Ejemplo: “Conviene restringir el acceso de nuestros datos sensibles a los desconocidos”
Difusión de rumores	Difusión de chismes a través del correo electrónico, mensajes de texto o sitios de redes sociales.
Etiquetar	Con el etiquetado o con los “rótulos” escribimos el nombre a las personas que figuran en alguna fotografía que se está dando a conocer en alguna red social o medio digital. Aunque puede ser un acto inocente también puede darse un etiquetado agresivo y no autorizado con el fin de acosar o burlarse de alguien. Ejemplo: “No debemos etiquetar a alguien en situación comprometida, como estando borracho o dormido”.
Exabyte (EB)	Unidad de datos equivalente a 10^3 (1.024) petabytes o 10^{18} bytes.
Geolocalización	Ubicación exacta y precisa proporcionada por el dispositivo (móvil, ipad…) acerca del lugar en el que se encuentra el usuario. Puede ser peligroso indicar dónde se encuentra uno en todo momento, especialmente si alguien lo está acosando o persiguiendo de algún modo.

Grooming.	Vocablo proveniente del inglés y que alude al “acicalamiento”, estableciéndose una analogía con la seducción que trata de producir el adulto que lo comete, para acercarse a un niño empleando los medios digitales y muchas veces haciéndose pasar por quien no es. El “Groomer” es el que comete este delito. Ejemplo: “El groomer le pidió al niño fotos o que se conectara a la webcam”.
Lingüística computacional	En las ciencias de la computación, la lingüística computacional estudia el idioma hablado para convertirlo en datos estructurados mediante herramientas computacionales.
Meme	Se refiere a una idea o imagen que se torna viral muy rápidamente en internet. Aunque con frecuencia es humorística, también puede contener algún contenido discriminatorio con carácter de burla y fines de acoso. Muy habitual en Twitter. Identidad de la persona dentro de las redes sociales, y que incluye información privada y datos sensibles acerca de su domicilio, gustos particulares, etc. Con el fin de evitar el acoso digital es conveniente limitar el acceso a nuestros familiares y amigos directos, configurando la privacidad. Ejemplo: “El acoso sexual se produce cuando el adulto crea perfiles falsos”.

NoSQL	Son sistemas de gestión de bases de datos y estructuras que permiten almacenar información cuando las bases de datos relacionales dan problemas.
Photoshop	La manipulación de imágenes digitales donde el sujeto principal se encuentre en una situación comprometida o embarazosa.
Predictive Analytics	El análisis predictivo es la utilización de datos para predecir tendencias o eventos futuros.
Sentiment Analysis	El análisis de sentimiento son técnicas usadas para identificar el sentimiento de un individuo sobre determinada cuestión. Hay muchos términos que surgen a cada momento, muchas veces creados por proveedores de herramientas y analistas de consultorías para intentar ofrecer un nuevo servicio.
Weka	Es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, clasificación, regresión, agrupación, extracción de reglas de asociación y visualización.

Resumen.

La presente investigación describe el proceso de un modelo predictivo, usado para categorizar los vocablos de Ciberbullying en sitios de Internet. El proceso se lleva a cabo, extrayendo las direcciones de las páginas Web, las cuales fueron localizadas a través de un Crawler, posteriormente se aplicó procesamiento de lenguaje natural, técnicas de la Web semántica, con el propósito de establecer una ontología semántica que ayude en la construcción del conjunto de datos (dataset) que será utilizado en las pruebas de predicción. El proceso de construcción del dataset, fue fundamentado a través de la herramienta ADVI (Castillo Zuñiga, I., Luna Rosas, F., Muñoz Arteaga, J., Lopez Veyna, 2016), la cual lleva a cabo la transformación de datos semi-estructurados a datos estructurados, finalmente generado en un archivo .csv. Para la detección del vocabulario de Ciberbullying, fueron seleccionados distintos algoritmos de aprendizaje de máquina (machine learning), en específico aprendizaje supervisado, Knn – vecinos más cercanos, Naive Bayes, Random forest, Arboles de Decisión, AdaBoost, C 4.5.

El modelo propuesto, demuestra que el mejor algoritmo para clasificar el contenido de sitios Web de la temática de Ciberbullying, son los Bosques Aleatorios, con una precisión del 94%. Por lo que podemos decir que es posible realizar dicha clasificación con un alto grado de confiabilidad.

Palabras Claves: Ciberbullying, Aprendizaje de Máquina, Aprendizaje Supervisado, Minería Web, Minería de texto, Big Data Analytics.

.

Abstract

The present investigation describes the process of a predictive model, used to categorize the words of Cyberbullying in Internet sites. The process is carried out, extracting the addresses from the Web pages, which were located through a Crawler, later applied natural language processing, Semantic Web techniques, to establish a semantic ontology that helps in the construction of the data set (dataset) that will be used in the prediction tests. The dataset construction process was based on the ADVI tool (Castillo-Zúñiga, et al., 2016), which carries out the transformation of semi-structured data into structured data, finally derived in a .csv file. For the detection of the Cyberbullying vocabulary, different machine learning algorithms were selected, in specifically supervised learning, Knn - neighbor's closets, Naive Bayes, Random forest, Decision Trees, AdaBoost, C 4.5.

The proposed model demonstrates that the best algorithm to classify the content of Web sites of the Cyberbullying theme is Random Forests, with an accuracy of 94%. So, we can say that it is possible to make such a classification with a high degree of reliability.

Keywords: Cyberbullying, Machine Learning, Supervised Learning, Web Mining, Text Mining, Big Data Analytics.

Agradecimiento.

Hoy culmina una etapa más de mi vida y quiero agradecer mis triunfos:

A los Amigos quienes incondicionalmente están presentes en mi vida, gracias a aquellos que con su respeto y decencia estuvieron presentes en el desarrollo de la tesis realizando aportes a esta, gracias a todos.

A mi esposa e hijos, mis sinceros agradecimientos por su apoyo incondicional y por el tiempo que me regalaron para realizar este documento.

A mis padres (q.e.p.d.), por su amor y a quienes les debo lo que soy, a mis hijos, por el amor, la alegría apoyo que me brindaron e hicieron posible estos logros. A mis familiares y demás por su apoyo ilimitado.

A todos los docentes y directivos de la UNIVERSIDAD CUAUHTEMOC, por la paciencia, enseñanza y dedicación brindada y en especial a nuestro director de tesis quien hizo que el proyecto se llevara a cabo, el Dr. Iván Castillo Zúñiga.

Finalmente, a mis compañeros de trabajo y estudio en especial a José, Israel y Alejandro, quienes hicieron que esta maestría pareciera algo más sencillo de lo que yo imaginaba y además me permitieron entrar en sus vidas.

Dedicatoria.

Agradecerle a DIOS que ha sido el forjador de mi camino y quien guía el destino de mi vida.

Agradezco a mi familia por apoyarme en cada decisión en mi vida, por su amor y bondad, hago presente mi gran afecto hacia ustedes, mi maravillosa familia.

Introducción.

En la era digital los datos tienen un crecimiento de manera exponencial, pero hay una cuestión, y esta es: ¿Cómo vamos a analizar estos datos? Hay algunas estadísticas de los datos realizada por el IDC (Inter nacional Data Corporation) (Requena-Mesa, 2018), el cual cuantifica y pronostíca la cantidad de datos producidos anualmente. Este estudio nos indica que en año 2013 habían 4.4 Zettabytes (4.4 trillones de gigabytes) y que para el 2020 llegaremos 44 Zettabytes (44 trillones de gigabytes), lo que señala que cada dos años su valor se duplica, lo que nos muestra que cada persona creara 1.7 MB cada segundo en información, también las estadísticas nos presentan que menos del 0.5 % de los datos mundiales se están analizando.

Con este crecimiento tan acelerado de los datos, aparecen infinidad de temas en la Web para consultar, y como caso de estudio en esta investigación, la relevancia es el enfoque que se le da a los vocablos tocantes a la temática de Ciberbullying.

En Colombia encontramos unas investigaciones dirigidas al Ciberbullying, realizando estudios en colegios para determinar en qué edades se presenta este tipo de acoso, que elementos tecnológicos se emplean, cual es el género que más es atacado y cuáles son las consecuencias que genera este tipo de acoso cibernético. Considerando este tipo de crimen cibernético, se plantea en este documento, las herramientas que se pueden emplear de Big Data analytics para poder ayudar a neutralizar este tipo crimen digital.

Los motivos por las cuales se inició este estudio, radican en la inquietud de explorar los beneficios que se pueden obtener de la información que circula en la Web, con el propósito de generar información fiable que sea útil en la toma de decisiones, para activar medidas de seguridad y prevención. La idea de la investigación, se centra en que es posible clasificar páginas Web, detectar vocabulario, y obtener conocimiento de Ciberbullying a partir del análisis de Páginas Web con técnicas de Big Data Analytics, Procesamiento de Lenguaje Natural, Web Semántica, y Aprendizaje Supervisado.

Capítulo I. Introducción.

1.1. Planteamiento del problema.

1.1.1 Contextualización del Ciberbullying.

Los delitos cibernéticos como es el acoso cibernético (Ciberbullying) a menudo ocurren en los sitios de redes sociales más populares como es el twitter. La práctica del asedio cibernético en adolescentes puede causar depresión, pensamientos asesinos o suicidas y necesita una acción preventiva para que no dañe a la víctima. El ciber-acoso o ciberacoso es un acto de atacar, humillar o dañar a otros intencional y repetidamente en los medios sociales, mensajes u otros medios de comunicación en línea. Al clasificar los tweets de datos para detectar el abuso cibernético en los sitios de redes sociales de Twitter, los datos se clasifican en dos clases: "Bullying" y "Not Bullying". La clase "Bullying" es una clase de tweets que contiene acciones de ciberacoso, mientras que la clase "Not Bullying" es una clase de tweets que no es una acción de ciberacoso (Anggraini, Sucipto, & Indriati, 2018).

Las investigaciones actuales han propuesto que el acoso sexual es uno de los problemas sociales más graves. La percepción del acoso sexual varía de acuerdo con algunos de los factores: género, contexto, y lado de la pareja. El rápido desarrollo de dispositivos y aplicaciones electrónicas para la comunicación ha cambiado recientemente la forma en que los adolescentes socializan. El uso de estos métodos de comunicación es prácticamente universal entre los jóvenes de los países tecnológicamente desarrollados. A pesar de las grandes ventajas que ofrecen estas herramientas para establecer nuevas amistades o mantener el contacto con la familia o amigos, también presentan algunos peligros potenciales significativos. A veces los adolescentes se

aprovechan del anonimato que ofrecen estos medios para dañar, ofender o herir intencionalmente a otros adolescentes. El término ciberagresión se utiliza típicamente para referirse a este tipo de comportamiento (Corcoran, Guckin, & Prentice, 2015).

La agresión a través de dispositivos electrónicos puede ser aún más grave que la agresión cara a cara. Puede ocurrir a cualquier hora del día y puede propagarse inmediatamente a muchas personas y, dado que ocurre a distancia, los perpetradores no ven los efectos negativos de la agresión en la víctima, lo que dificulta la empatía y promueve la recurrencia (Álvarez-garcía, Barreiro-collazo, Nuñez, & Dobarro, 2015).

La ciberagresión entre adolescentes ocurre en muchos países y constituye un problema internacional según (Álvarez-garcía et al., 2015). Sin embargo, el porcentaje estimado de adolescentes involucrados en tales actos varía según la definición de ciberagresión, tomando en cuenta la muestra y la metodología utilizada para el análisis. Los chicos presentan una mayor tendencia que las chicas en enviar contenido sexual no deseado, y en grabar y difundir imágenes degradantes sobre la víctima.

Además de estas consecuencias psicológicas y de comportamiento, la ciberagresión también puede tener consecuencias legales. En España, el Código Penal tipifica como delitos las siguientes conductas: amenazas, coacción, calumnia, engaño por teléfono o Internet de un menor de 16 años para obtener material pornográfico en el que aparezca un menor, o delitos contra la intimidad. Menores infractores menores de 14 años de edad no pueden ser considerados penalmente responsables, y la Administración,

a través de los Servicios Sociales, es promover medidas de control, reeducar y proteger en casos de abandono familiar. El menor de edad la familia asumirá la responsabilidad civil que pueda derivarse de la responsabilidad actuar (González-Garcia, 2016).

1.1.2 Definición del problema.

Actualmente la mayoría de los estudios de vocabulario son enfocados al análisis de sentimientos y realmente no se cuentan con estudios que identifiquen en los distintos sitios web el vocabulario referente al Ciberbullying. La literatura muestra escasos estudios que apliquen técnicas de aprendizaje de máquina (machine learning) para la detección de vocabularios.

El problema que aborda la investigación es, “El análisis, clasificación y predicción del vocabulario de Ciberbullying a partir de datos obtenidos de páginas de Internet”.

Problemática de manera específica:

1. Dificultad para transformar la información proveniente de las páginas Web del Ciberbullying a datos estructurados (dataset), debido a que la información se encuentra mezclada con código HTML, PHP, además contiene errores de sintaxis, polimorfismo en las palabras y errores ortográficos.
2. Dificultad para obtener valor agregado en la información (que permita sustentar la toma de decisiones), debido a la complejidad de las técnicas de aprendizaje supervisado que son implementadas.

3. Dificultad para llevar a cabo la clasificación de las páginas Web, ya que es necesaria la implementan de algoritmos de aprendizaje de máquina (machine learning).
4. Dificultad para predecir el vocabulario de Ciberbullying.

1.1.3 Preguntas de investigación.

1. ¿Es posible utilizar los algoritmos de machine learning como Arboles de decisión, Método K-vecinos más cercanos, Naive Bayes, Random Forest, Máquinas de Soporte Vectorial (SVM), Regresión Logística Lineal, Redes Neuronales y Adaboost para la detección de vocabulario referente a Ciberbullying?
2. Al ser evaluado el dataset de Ciberbullying con las distintas técnicas de machine learning, ¿Cuales fueron las técnicas más adecuadas en la detección de vocablos de Ciberbullying?.

1.2. Justificación.

Las tecnologías actuales de Big Data surgieron con el objetivo de cubrir unas necesidades no satisfechas para encontrar tendencias globales de los datos, estos datos son poco estructurados y en cantidades gigantescas, los cuales son difíciles de explotar con métodos convencionales. El Big Data tiene unas características muy concretas y que son: Volumen (Tamaño de los Datos), Velocidad (Rapidez con que llegan los datos) y Variedad (Los datos son estructurados, Semi-estructurados y No estructurados). Esta investigación nos va a permitir conocer los síntomas producidos por el uso frecuente y

los diferentes tipos de Ciberbullying, teniendo como fuente las diferentes redes sociales que hacen propensas a las diferentes personas a ser víctimas de ataques virtuales, el uso excesivo de las redes sociales puede generar riesgos de soledad y cambios en la forma de comunicarse afectivamente con su círculo familiar y de amigos afectando la autoestima de las personas.

.

1.2.1. Conveniencia.

Una de las características del Ciberbullying es que los ataques que se realizan a los usuarios son ataques individuales y agresivos, se ha comprobado mediante diversas investigaciones que los estudiantes que usan el Internet menos de una hora tienen menos riesgo de ser victimados que los que usan el Internet durante más tiempo. Los comentarios ofensivos y burlas lesionan la estabilidad emocional y causan daño a la autoestima de la población preadolescente y adolescente llevándolos a producir cuadros depresivos. En la red se ofrecen algunas orientaciones para evitar el Ciberbullying como es no contestar las amenazas, bloquear a los abusadores, no dar información personal, no publicar fotos comprometedoras, denunciar estos casos.

1.2.2. Relevancia social en Colombia.

En Colombia se han definido desde el 2011 varios documentos CONPES enfocados a temas de seguridad y hacia el año 2016 se definió uno sobre seguridad digital, todo esto corroborar la importancia que se da cada vez más a los temas de seguridad digital, por ende, es de gran importancia social poder contar con distintas herramientas que permitan la defensa en cuanto a ataques de seguridad digital. El

análisis de vocabulario a partir de técnicas de Big Data son algunas de las herramientas que pueden ser utilizadas para estos fines.

1.2.3. Implicaciones educativas.

Podemos considerar como una violencia injustificada al Acoso Cibernético, donde impera un claro desequilibrio de autoridad entre el culpable y la víctima, sostenida en el tiempo habitualmente bajo el código del silencio. Las nuevas tecnologías proporcionan a los jóvenes muchas ayudas para el acceso a la información y que tenga mucha incidencia en sus relaciones interpersonales tanto en su casa como con el grupo de compañeros y amigos, como consecuencia podemos observar conductas agresivas, intimidatorias y de acoso, si estos tienen una utilización inadecuada de las nuevas tecnologías.

1.2.4. Relevancia teórica.

Los individuos que sufren de Acoso Cibernético se convierten en personas acomplejadas, retraídas, algunos de ellos no quieren asistir a sus colegios por temor a las amenazas a que son sometidos o a las burlas. Debido al hostigamiento que son sometidos por las diferentes redes sociales, ellos creen que están a la moda y conocen las últimas noticias que suceden en el mundo. Estas redes sociales permiten a los usuarios el acercamiento con los conocidos que no viven cerca de ellos o tener un acercamiento por los que debido a sus ocupaciones diarias no es posible hacerlo de manera personal. Cualquier persona que ingrese a una red social puede acceder y este puede aceptar a un individuo conocido o no, y puede convertirse en una víctima ya que

no cuenta con una persona responsable que lo supervise y lo guie en el momento de estar navegando.

1.2.5. Utilidad metodológica.

Este estudio con la información consignada puede convertirse en un instrumento para analizar, recoger información referente al Acoso Cibernético aplicando diferentes técnicas de Análisis de Datos de Big Data.

1.3. Objetivos.

1.3.1 Objetivo general.

Aplicar técnicas de Machine Learning en el análisis de datos para predecir vocabulario de Ciberbullying, usando páginas de Internet descargadas por una aplicación para su posterior análisis.

1.3.2 Objetivos específicos.

1. Utilizar distintas técnicas de aprendizaje de máquina para la detección de vocabulario de Ciberbullying.
2. Encontrar el algoritmo adecuado para clasificar de manera confiable la detección de vocabulario de Ciberbullying.

1.4. Hipótesis.

Es posible predecir el vocabulario del Ciberbullying, clasificar sitios Web y obtener un valor agregado en Internet sobre las páginas que circulan en este, a través de la aplicación de técnicas de Análisis de Big Data, Procesamiento de Lenguaje Natural, Web Semántica y Aprendizaje de Máquina (Aprendizaje Supervisado).

1.5. Breve descripción de la organización de la tesis.

La estructura de este trabajo ésta formado por cinco capítulos, las referencias bibliográficas y sus anexos.

El capítulo 1, Introducción, tiene como objetivo proveer al lector del contexto de la investigación, resumiendo el problema de estudio atacado, los aportes previos al tema investigado, la definición de conceptos claves y todo aquello que ayude al entorno mencionado en los párrafos respectivos. Se define la justificación del trabajo el cual es la conveniencia de la investigación y cuáles son los beneficios que se esperan con este conocimiento obtenido. La viabilidad es otro punto tratado en este capítulo son los recursos disponibles para llevar a cabo la investigación, como son los humanos, materiales y financieros. Por último, tenemos la hipótesis que nos sirve de forma alternativa para responder un problema con base científica.

El capítulo 2, Se desarrolla el marco teórico que nos permite la interpretación de resultados, tenemos el estado del arte donde se describen los antecedentes de la investigación. Se establecen las bases teóricas en las que está sustentado el trabajo de tesis, se enfoca en conceptos relacionados con Big Data, Web Semántica, Procesamiento de Lenguaje Natural, Pre-Procesamiento de Datos y Aprendizaje de Máquina. Por último, se aborda de manera puntual el tema del Ciberbullying, el cual es el caso de estudio en la investigación.

El capítulo 3, Se tocan temas como el objetivo general y los específicos, materiales y métodos, presenta una descripción detallada de la propuesta de solución a la problemática planteada y a los objetivos establecidos al inicio de la investigación.

El capítulo 4, Pruebas y resultados, se muestran y explican los productos obtenidos tras la evaluación/confirmación de la propuesta de solución desarrollada.

El capítulo 5, Conclusiones, se puntualizan las ventajas y desventajas de la investigación, considerando la problemática y los objetivos establecidos al inicio del estudio. Además, se presentan las posibilidades de trabajo futuro. Las referencias bibliográficas se muestran en formato APA en orden alfabético y para finalizar se incluyen los anexos.

Capítulo II. Estado del arte.

2.1. Ideas, procedimientos y teorías relacionadas al problema de investigación.

2.1.1. Grandes datos en la Web.

Desde que el hombre ha tenido la necesidad de guardar información escribiéndola en cuevas como la de Altamira en España, posteriormente se desarrolló la escritura cuneiforme 3400 años antes de cristo hasta que Johann Gutenberg invento la imprenta en 1440 y posteriormente se desarrollan las computadoras en el siglo XX; los seres humanos hemos generado una inmensidad de información y por consiguiente creamos la necesidad de resguardarla, administrarla y posteriormente sacarle provecho a estas grandes cantidades de datos. Es aquí con la aparición de los computadores que realmente necesitamos aprovechar estos grandes datos. Pero analizar grandes datos y encontrar patrones es un proceso en constante evolución ya que (Martínez-Gamboa, 2016) lo plantea: desde el punto de vista literario, diversos especialistas provenientes de las ciencias de la computación, la lingüística del corpus y la lingüística computacional han intentado llevar a cabo tareas de recuperación de información, análisis, interpretación, detección de patrones y detección de tendencias no sólo en los textos digitales de orden general, sino que también, e incluso de manera más intensa, en los textos literarios. Dicho en simple, cuando se dispone de todas las palabras de un texto en formato digital, es posible aplicar sobre lo textos rutinas de análisis automático que son capaces de detectar patrones que de manera manual sería imposible de encontrar.

También es importante observar desde el punto de vista empresarial que Big Data tiene la capacidad de cambiar la naturaleza de un negocio. De hecho, hay muchas empresas cuya única existencia se basa en su capacidad para generar información que solo Big Data puede ofrecer. Las empresas deben comprender que Big Data no se trata sólo de tecnología, sino también de cómo estas pueden impulsar a una organización hacia adelante. (Erl, Khattak, & Buhler, 2014).

Según (Mahmood, 2016), Big Data, ya no está restringido a salidas visuales y bases de datos clásicas; también incluye datos altamente desestructurados en forma de documentos de texto, páginas Web, fotos, datos espaciales y multimedia, información gráfica, comentarios en redes sociales y opiniones públicas. Dado que Big Data se caracteriza por tamaños de muestra demasiado grandes y bastante heterogéneos, y debido a la acumulación de ruido en los datos, los enfoques tradicionales de gestión de datos, visualización y análisis no son de aplicación satisfactoria. Por lo tanto, hay una necesidad urgente de nuevas herramientas, mejores marcos y metodologías viables para que dichos datos sean adecuadamente categorizados, segmentados lógicamente, analizados eficientemente y manejados de manera segura. Es aquí donde se convierte fuerte y fundamental Big data.

2.1.2. Ciberbullying.

Hacia el año 2014 Vandebosch recopila un compendio de autores y trata de definir el acoso cibernético, por definición, es un tipo de acoso escolar.

La intimidación es un comportamiento agresivo o «daño» intencional que se lleva a cabo «repetidamente y con el tiempo» y en una relación interpersonal caracterizada por un desequilibrio de poder. Otra diferencia importante es la falta de señales físicas y sociales en la comunicación en línea.

Esto evita que el acosador se enfrente con las consecuencias de los acosos y también podría conducir a una mala interpretación de los mensajes como acoso cibernético cuando, de hecho, no estaban destinados a ser.

En la mayoría de los casos, las víctimas del acoso tradicional están a salvo en el hogar. Este ya no es el caso del ciber-acoso. (Vandebosch, 2014)

2.1.3. Aprendizaje Supervisado.

Para (Lantz, 2013), este tipo de aprendizaje proporciona un conjunto de herramientas que utilizan los ordenadores para transformar los datos en conocimiento práctico. Se parte en dos áreas principales el Machine Learning: aprendizaje supervisado y aprendizaje no supervisado. Aunque pueda parecer que el primero permite buscar patrones en datos históricos relacionando los campos con un campo especial con intervención humana y la segunda usa datos históricos que

no están etiquetados para tratar de organizarlos, estos dos conceptos tienen más que ver con qué queremos hacer con los datos.

Lo que deseamos es formular algún tipo de regla o correspondencia que nos permita dar la respuesta para todos los objetos que se nos presenten. Por ejemplo, podemos contar con información de casas, y los precios correspondientes de cada una. A partir de esto, queremos aprender alguna regla de correspondencia entre casas y sus precios, que nos permita predecir el costo cuando tengamos información de nuevas casas. El entrenamiento de un algoritmo se realiza con un grupo de objetos con sus respuestas que deseamos estimar (precios).

Las máquinas de aprendizaje son utilizadas en lo siguiente:

- Predecir los resultados de las elecciones.
- Identificar y filtrar los mensajes de spam del correo electrónico.
- Prever la actividad delictiva.
- Automatizar los semáforos de acuerdo con las condiciones de la carretera.
- Producir estimaciones financieras de tormentas y desastres naturales.
- Examinar la rotación de clientes.
- Crear aviones de auto-pilotaje y coches de auto-conducción.
- Identificar a las personas con capacidad para donar.
- Publicidad dirigida a tipos específicos de consumidores.

Un algoritmo en máquina de aprendizaje toma los datos e identifica patrones que pueden ser utilizados para la acción. En algunos casos, los resultados son tan exitosos que parecen llegar a un estatus casi legendaria. Independientemente de si el alumno es un ser humano o una máquina, el proceso básico de aprendizaje es similar, lo podemos dividir en tres componentes de la siguiente manera:

- **Entrada de datos**: Utiliza la observación, el almacenamiento de memoria y el recuerdo para proporcionar una base factual para el razonamiento adicional.
- **Abstracción**: Implica la traducción de los datos en representaciones más amplias.
- **Generalización**: Utiliza datos abstractos para formar una base para la acción.

Una mejor estrategia es dedicar tiempo a la gestión selectiva de un conjunto más pequeño de ideas clave. Las estrategias de aprendizaje comúnmente utilizadas para crear un esquema o un mapa conceptual son similares a la forma en que una máquina realiza la abstracción del conocimiento. La generalización requiere una amplia gama de datos abstractos, así como una comprensión de más alto nivel de cómo aplicar dichos conocimientos a temas imprevistos.

Porque a los modelos predictivos se les da una instrucción clara sobre lo que necesitan para aprender y cómo se pretende que aprendan, se conoce el

proceso de formación de un modelo predictivo como aprendizaje supervisado. Específicamente, dado un conjunto de datos, el aprendizaje intenta optimizar una función para encontrar la combinación de valores de característica que dan como resultado la salida de destino. Es fácil pensar en los usos potenciales para un clasificador. La característica objetivo a predecir es una característica categórica conocida como la clase y está dividida en categorías llamadas niveles.

Los estudiantes supervisados también pueden ser utilizados para predecir datos numéricos como el ingreso, valores de laboratorio, resultados de pruebas o conteos de artículos. Para predecir tales valores numéricos, una forma común de predicción numérica ajusta los modelos de regresión lineal a los datos de entrada. Aunque los modelos de regresión no son el único tipo de modelos numéricos, sí lo son con mucho, el más utilizado.

Funcionamiento y tipos.

La primera modalidad de aprendizaje que tiene el machine learning es la de aprendizaje supervisado. Usándola, se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas. Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones. Existen, a su vez, dos tipos de aprendizaje supervisado:

- **Regresión:** tiene como resultado un número específico. Si las etiquetas suelen ser un valor numérico, mediante las variables de las características, se pueden obtener dígitos como dato resultante.
- **Clasificación:** en este tipo, el algoritmo encuentra diferentes patrones y tiene por objetivo clasificar los elementos en diferentes grupos**.** El algoritmo no está en capacidad de determinar a qué grupo pertenece un valor o cuál es el resultado de una operación. Solamente logra relacionar características con etiquetas y así obtener un resultado.

2.1.4 Técnicas de Máquinas de Aprendizaje (Machine Learning).

Según (Vandebosch et al., 2016) el aprendizaje automático o aprendizaje de máquinas es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. El aprendizaje automático es un proceso durante el cual una máquina aprende sobre patrones significativos en un conjunto dado de datos de entrada.

Los algoritmos de máquinas de aprendizaje se pueden dividir en dos grupos principales: supervisados que se utilizan para construir modelos predictivos, y estudiantes no supervisados que se utilizan para construir modelos descriptivos. Un modelo predictivo se utiliza para tareas que implican, como su nombre indica, la predicción de un valor utilizando otros valores del set de datos. A pesar del uso

común de la palabra «predicción» para dar a entender que los modelos predictivos de pronóstico no tienen por qué prever necesariamente los acontecimientos futuros.

Aplicaciones del Aprendizaje Automático.

- Un servicio de correo electrónico filtra el correo no deseado.
- Una cámara digital detecta rostros humanos.
- Los asistentes digitales personales (PDA) detectan comandos de voz.
- El software de reconocimiento óptico de caracteres (OCR) reconoce los caracteres de una imagen.
- Las pasarelas de pago en línea detectan el fraude de tarjetas de crédito.
- Los sitios web muestran anuncios personalizados a un usuario.

Además de las aplicaciones mencionadas anteriormente, el ML también se utiliza en varios otros dominios, como el diagnóstico médico, la exploración espacial, la conservación de la vida silvestre, el análisis de sentimientos, etc.

En términos sencillos, se puede pensar que el NM convierte las experiencias pasadas en experiencia. Consideremos la tarea de reconocer correos electrónicos no deseados y etiquetarlos. Ahora, cuando aparece un nuevo correo electrónico, la máquina busca palabras sospechosas identificadas en el conjunto de correos electrónicos no deseados anteriores para identificar si el nuevo correo electrónico es un correo no deseado. De esta manera, la máquina podrá etiquetar nuevos correos electrónicos de forma correcta y automática.

Tipos de Aprendizaje de Máquina (Machine Learning).

En general, hay tres tipos de técnicas de Machine Learning:

- **Aprendizaje supervisado**: este tipo de algoritmo de aprendizaje, consiste en un conjunto de datos de entrenamiento que tiene pares que consisten en valores de entrada y valores de salida deseados. En tales algoritmos, se genera una función de mapeo que procesa los valores de entrada en el conjunto de datos de entrenamiento a sus valores de salida correspondientes.

- **Aprendizaje no supervisado**: este tipo de algoritmo de aprendizaje, consiste en un conjunto de datos de entrenamiento que sólo tiene valores de entrada, no valores de salida. En dichos algoritmos de aprendizaje, el aprendizaje se produce únicamente en función de la estructura de los datos. El aprendizaje no supervisado representa el concepto de computación que se enseña a sí mismos. Se utiliza principalmente para resolver tareas de clustering.

- **Aprendizaje de refuerzo:** este tipo de algoritmo de aprendizaje interactúa con el entorno al producir acciones y descubrir errores o éxitos. Este tipo de aprendizaje permite a las máquinas determinar automáticamente el comportamiento ideal dentro de un entorno específico. Se requiere una

retroalimentación simple para que la máquina sepa qué acción es la mejor; Esto se conoce como la señal de refuerzo**.**

2.2 Descripción de trabajos relacionados.

2.2.1. Cyberbullying Detection Modelling at Twitter Social Networking.

(Anggraini et al., 2018) afirman que este artículo trata sobre un delito cibernético ocurrido en una de las redes sociales más populares como es el Twiter, informando sobre el acoso cibernético como un delito y sus respectivas consecuencias.

Para prevenir el ciberacoso se puede hacer un modelo de minería de texto para clasificar los tweets en Twitter en dos clases, la clase de acoso y la clase de no acoso. En esta investigación utilizamos Naïve Bayes Classifier con cinco etapas de pre-procesamiento: reemplazar tokens, transformar case, tokenization, filtrar stopwords y n-grams. Para evaluar el rendimiento del modelo se utiliza una tabla de matriz de confusión. El modelo en la fase de validación cruzada de 10 veces funciona bien con 77,88% de precisión, 94,75% de retirada y 82,50% de precisión con +/-5,12% de desviación estándar.

A partir de los resultados realizado por los investigadores, se puede concluir que la detección del acoso cibernético en las redes sociales de Twitter se puede hacer con varias técnicas. En segundo lugar, el proceso de selección de datos, la limpieza de datos y el pre-procesamiento se llevan a cabo para preparar los datos

en el proceso de minería. Los resultados del proceso de modelado de detección cibernética en el proceso de validación cruzada de 10 veces tienen una precisión promedio de 77.88%, un recordatorio de 94.75% y una precisión de 82.50% con una desviación estándar de precisión de +/- 5.12% que muestra que el modelo es un modelo estable.

2.2.2. Validity and reliability of the Cyber-aggression Questionnaire for Adolescents (CYBA).

El artículo de (Álvarez-García, Barreiro-Collazo, Núñez, & Dobarro, 2015) plantea que la ciber delincuencia es un problema creciente y preocupante, especialmente cuando se trata de menores. La ciberagresión, en particular entre los adolescentes, puede tener consecuencias legales y psicológicas negativas para las personas involucradas. Los objetivos de este documento son diseñar un nuevo informe y un Cuestionario de Ciber-Agresión para Adolescentes, para evaluar los resultados hasta qué punto el encuestado realiza agresiones a través de un teléfono móvil o Internet y analizar la validez y fiabilidad factorial y de criterio de sus puntuaciones en una muestra de adolescentes de Asturias, España.

El CYBA administró a 3.148 jóvenes de entre 12 y 18 años de edad, junto con tres autoinformes para medir la agresión en la escuela, la impulsividad y la empatía. En cuanto a la validez de los criterios, la puntuación del CYBA se correlaciona positivamente con la agresión en la escuela y la impulsividad y

negativamente con la empatía. La confiabilidad de los puntajes en cada ítem y factor del CYBA son adecuados.

El rápido desarrollo de dispositivos y aplicaciones electrónicas para la comunicación ha cambiado recientemente la forma en que los adolescentes socializan. El uso de estos métodos de comunicación es prácticamente universal entre los jóvenes de los países tecnológicamente desarrollados. Estos porcentajes son similares a los de otros países europeos, como Francia, Irlanda o Suecia. A pesar de las grandes ventajas que ofrecen estas herramientas para establecer nuevas amistades o mantener el contacto con la familia o amigos (Corcoran et al., 2015), también presentan algunos peligros potenciales significativos. El término ciberagresión se utiliza típicamente para referirse a este tipo de comportamiento. La agresión a través de dispositivos electrónicos puede ser aún más grave que una agresión cara a cara (Álvarez-García∗ et al., 2015).

2.2.3. Simbiosis vital para describir el Ciberbullying en Colombia.

El Ciberbullyng es un tema de gran interés en Colombia así lo plantea el artículo (Rincón Rueda, Alberto Isaac; Ávila Díaz, 2014), que plantea que por el excesivo contacto que tienen los preadolescentes y los adolescentes con la tecnología, lo que los distancia cada vez más del mundo real. Por eso, los objetivos de este escrito son saber por qué el Ciberbullying ha crecido en Colombia y reflexionar sobre las posibles soluciones que podrían minimizar el impacto psicológico ocasionado. Se identificaron las posibles soluciones que han dado los

organismos de control y las instituciones educativas para combatir la intimidación en línea que cada día cobra mayor fuerza en las redes sociales.

Para analizar este problema, es necesario mencionar sus causas y una de ellas es su permanencia en las redes sociales, las cuales se entienden como «la existencia de las redes humanas determinadas por las tecnologías y la ciencia en las que el ser se ha constituido como un número o como un código de identificación». Se ve entonces, cómo, sin ninguna oposición, ellos —los jóvenes que viven su preadolescencia y adolescencia— aceptan formar parte del consumismo que traen las tecnologías, las cuales no hacen distinción de clase, raza, sexo o condición humana.

La realidad del Ciberbullying consiste en hacer uso de las tics, principalmente, para ejercer el acoso psicológico entre iguales, y se da, en su gran mayoría, en las instituciones educativas cuando los preadolescentes y adolescentes se acusan con chantajes e insultos, entre otros, que los van perjudicando. Ello permitió identificar las posibles soluciones que se están dando para minimizar el impacto psicológico que puede ocasionar el acoso virtual en el grupo comprendido entre los 10 y 20 años de edad, aproximadamente.

Particularmente, el Ciberbullying se ha dividido en siete subtipos: 1) mensajes de textos recibidos en el teléfono móvil; 2) fotografías o videos realizados con las cámaras de los móviles y posteriormente utilizados para amenazar a la víctima; 3) llamadas acosadoras al teléfono móvil; 4) mensajes de correo electrónico

insultantes o amenazantes; 5) salas de chat en las que se arremete contra uno de los participantes o se le excluye socialmente; 6) acoso mediante programas de mensajería instantánea; y 7) páginas web donde se difama a la víctima al publicar información personal en la red o se hacen concursos en los que se ridiculiza a los demás.

La sociedad actual colombiana ha ido construyendo su existencia en torno a las tecnologías como medio para estar conectado con el mundo entero. Sin embargo, estas se han constituido en una forma de dominar a las personas. De la misma manera, llamamos la atención sobre el hecho de que las tecnologías se han trasformado en un medio esencial para generar más violencia contra los derechos humanos, especialmente de los preadolescentes y adolescentes, a tal punto de que las autoridades han intentado crear normas para el buen uso de las tecnologías y prever el matoneo y el acoso cibernético.

2.2.4. Weakly supervised Cyberbullying detection with participant‐vocabulary consistency.

El acoso en línea y el ciberacoso como lo plantea el artículo (Raisi & Huang, 2018), comenta que se están convirtiendo en graves amenazas sociales para la salud que dañan la vida de las personas. Para abordar la naturaleza evasiva del acoso cibernético con un mínimo de esfuerzo y costo, el algoritmo de aprendizaje sólo requiere una débil supervisión. Ilustramos las fortalezas y debilidades de él modelo mediante el análisis de las conversaciones y frases claves identificadas por PVC.

De acuerdo con stopbullying.gov, hay varias formas de acoso cibernético, pero no limitado al acoso, difusión de rumores, y la publicación de imágenes vergonzosas. Presentamos un método automatizado y basado en datos para identificación de acoso.

El análisis del acoso en línea requiere un enfoque multifacético, comprensión del idioma y de las estructuras sociales. En las complejidades que subyacen a estos comportamientos hacen que sean difícil de detectar para los enfoques computacionales estáticos. Sin embargo, para entrenar a un método de aprendizaje supervisado de la máquina, los datos de entrada etiquetados son necesario. La anotación de datos es un proceso costoso y que requiere mucho tiempo. La supervisión débil es que el algoritmo de aprendizaje debe ser capaz de encontrar patrones en los datos no etiquetados para integrarse con la débil supervisión.

Utilizamos este débil paradigma de supervisión para aliviar significativamente la necesidad de expertos humanos para realizar tediosas anotaciones de datos. Nuestra contribución tiene como objetivo mejorar la detección automatizada del ciberacoso. Por ejemplo, actualmente estamos desarrollando enfoques débilmente supervisados que consideran adicionalmente las características de la red o la secuencia de conversaciones. Lo que debemos hacer cuando se detecta el acoso cibernético es un importante punto de partida.

2.2.5. Evaluar el Bullying y el Ciberbullying validación española del EBIP-Q y del ECIP-Q.

El bullying es un fenómeno de agresión injustificada habla el artículo de los autores (Ortega Ruiz, Rey, & Casas, 2016), el cual menciona que actualmente sucede en dos formatos: cara a cara y como una conducta que se realiza a través de dispositivos digitales. Este trabajo presenta la validación del European Bullying Intervention Project Questionnaire y del European Cyberbullying Intervention Project Questionnaire, que evalúan la implicación en Bullying y en Ciberbullying, respectivamente. La realización de un modelo de ecuaciones estructurales ha evaluado la concurrencia y relaciones entre ambos fenómenos, encontrando la influencia del Bullying sobre el Ciberbullying, pero no al contrario.

Pensando en la necesidad de la intervención psicoeducativa y la práctica de la psicología escolar, disponer de ambos instrumentos homogéneos en cuanto a sus medidas para su uso en el contexto del centro y el aula escolar puede facilitar la rápida descripción de los niveles de prevalencia de ambos problemas, lo que puede ser muy útil a los profesionales de la psicología escolar y a los agentes del sistema educativo, en su trabajo de prevención tanto del Bullying como del Ciberbullying.

El segundo objetivo de este trabajo ha buscado descubrir la relación de influencia y solapamiento o tránsito de un fenómeno al otro y viceversa. Estos resultados apoyan la idea de que los principales roles de implicación se mantienen

cuando el Bullying pasa a realizarse y quizás a perpetrarse con el uso de medios tecnológicos de comunicación. Quizás el factor de dominio tecnológico parece que podría estar jugando contra el agresor tradicional, que sabiendo cómo controlar, intimidar y acosar a otro de forma directa, en el escenario virtual es menos hábil porque no necesariamente el agresor es más competente digitalmente que los demás. Lo que definitivamente se evidencia en nuestros resultados es que el agresor de Bullying tradicional puede volverse víctima a través de medios digitales.

Debemos tener en cuenta las limitaciones que presenta este trabajo, que son las propias de la mayor parte de los auto-informes, donde la deseabilidad social puede afectar a los resultados.

2.2.6. Bullying y Ciberbullying en Colombia; co-ocurrencia en adolescentes escolarizados.

(Herrera-López, Romera, & Ortega-Ruiz, 2017), plantea que el primer objetivo fue adaptar y comprobar las propiedades psicométricas de la escala de Bullying: European Bullying Intervention Project Questionnaire para Colombia, dada la necesidad de disponer de instrumentos validados y universalmente homologados. Participaron 1931 adolescentes colombianos, 53% chicas. Se realizaron análisis aplicados a los ítems, análisis factoriales confirmatorios y modelos de ecuaciones estructurales. La prevalencia de implicación en Bullying fue del 41.9% y para Ciberbullying del 18.7%. Un segundo objetivo buscó encontrar la prevalencia de Bullying y Ciberbullying en una muestra colombiana. El tercero fue analizar el grado de relación y coocurrencia de los fenómenos.

Para obtener la adaptación colombiana inicialmente los instrumentos fueron sometidos a una validez de contenido por medio del juicio de seis expertos, quienes valoraron los criterios de adecuación del vocabulario, claridad conceptual, coherencia y relevancia de cada ítem. Finalmente, se realizó una prueba piloto con 60 escolares para valorar el grado de comprensión de los ítems.

Así, para determinar el rol de víctima o cibervíctima se consideraron sujetos con calificaciones iguales o superiores a 2 en cualquiera de los ítems de victimización, y con puntaje igual o menor que 1 en todos los ítems de agresión. La implicación en el rol de agresor o ciberagresor se calculó considerando los sujetos con puntuaciones iguales o superiores a 2 en cualquiera de los ítems de agresión, y con puntaje igual o menor que 1 en todos los ítems de victimización.

Los hallazgos revisten importancia para el ámbito educativo, pues, además de ofrecer al contexto colombiano un instrumento de medida de calidad e internacionalmente homologado para beneficio de los estudios comparativos, de prevalencia e intervención, deja manifiesta la necesidad de intervenir estos fenómenos desde modelos ecológicos y globales que los asuman integralmente y que actúen sobre la convivencia directa pero también sobre la ciber-convivencia, sin que ello impida acciones específicas para cada uno por separado.

2.3. Análisis de trabajos relacionados.

La Tabla 1, presenta un análisis de las características de los trabajos relacionados al problema de investigación. Lo que permite identificar y comparar las técnicas y procedimientos que han sido utilizados en investigaciones similares. En el mismo sentido, se resaltan las diferencias de la propuesta actual contra los trabajos comparables.

Tabla 1. Características de trabajos relacionados para la detección de Ciberbullying

Trabajos Relacionados	Tipo de investigación		Orientación del estudio	Técnicas para detección de Cyberbullying																						Herramientas de Minería de datos
	Cuantitativa	Cualitativa			Conjunto de datos (dataset)									Aprendizaje supervisado												
				Fuente de datos (Corpus)	Fuente de datos (Dataset)	TF	TF - IDF	Stop	Sinonimo	Crawler	SFC	LDA-TOT	A-TOT	Bayes	SVM	Árboles	K-NN	Regresión Lineal	J48	Redes Neuronale	C5.0	Bosques Aleatorios				
Álvarez-García 2016	↑	↑	Cuestionario Cyberacoso	CYBA	CYBA		↑	↑						↑	↑		↑									Desarrollo Propio
Anggraini 2018	↑	↑	Prevención de acoso	Tweets de Twitter	Rastreo de Datos						↑				↑		↑	↑								Naïve Bayes
González-Garcia 2016	↑	↑	Cuestionario de Sexting	Software estadístico SAS	Chi-squared statistic			↑						↑	↑	↑	↑	↑	↑							Desarrollo Propio
Herrera-López 2017	↑	↑	Pruebas de contraste de proporciones	Escala EBIPQ y Análsis basados en la TRI	Escala EBIPQ y Análsis basados en el Modelo TRI									↑	↑		↑									Desarrollo Propio
Ortega Ruiz 2016	↑	↑	Cuestionario bullying y cyberbullying	Escala ECIPQ	Escala ECIPQ															↑	↑					AFC
Moohebat	↑	↑	Vocabulario	Revistas	Revistas			↑				↑	↑													Desarrollo Propio
Basher 2014	↑	↑	Vocabulario	Mensajes de Yahoo, AOL	Mensajes de Yahoo, AOL	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑		↑		↑				Desarrollo Propio

Capítulo III. Materiales y métodos.

3.1. Fundamento del dataset.

El corpus lingüístico se administra a través de ontologías semánticas formadas con metadatos, donde la clase representa el tema principal, la subclase a los subtemas y los atributos a cada palabra que hace referencia a cada característica, como se aprecia en la Figura. 1.

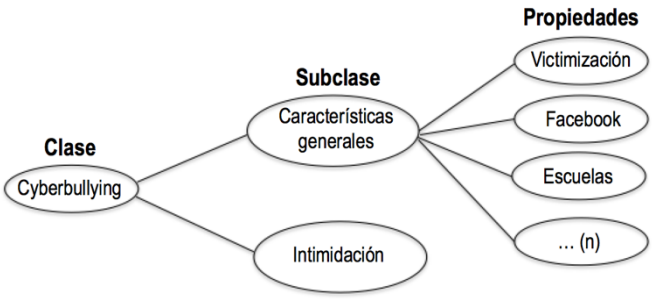


Figura 1. Modelo de ontologías orientado a objetos para Ciberbullying

En esta tesis se consideró el corpus lingüístico de Ciberbullying, el cual se describe en la Tabla 2. Es importante mencionar que el corpus lingüístico es establecido para el idioma español.

Tabla 2. Corpus lingüístico de Ciberbullying.

No.	Término	No.	Término	No.	Término	No.	Término
1.	Agresivo.	19.	Vergüenza.	37.	Primaria.	55.	Facebook.
2.	Aislamiento.	20.	Violencia.	38.	Secundaria.	56.	Fotografía.
3.	Amenaza.	21.	Abuso.	39.	Preparatoria.	57.	Grabación.
4.	Ansiedad.	22.	Cambios.	40.	Colegio.	58.	Internet.
5.	Ataque.	23.	Ciberacoso.	41.	Universidad.	59.	Mensajes.
6.	Autoestima.	24.	Confidencial.	42.	Alumnos.	60.	Móviles.
7.	Ciberbullying.	25.	Denigrante.	43.	Profesor.	61.	Página.
8.	Depresión.	26.	Divulgar.	44.	Acosador.	62.	Teléfono.
9.	Estrés.	27.	Emoción.	45.	Testigo.	63.	Texto.
10.	Hostigar.	28.	Espiar.	46.	Víctima.	64.	Video.
11.	Humillar.	29.	Falsedad.	47.	Cómplice.	65.	YouTube.
12.	Insultos.	30.	Humor.	48.	Adolecente.	66.	Twitter.
13.	Intimidar.	31.	Intención.	49.	Niño.	67.	Virtual.
14.	Manipular.	32.	Ira.	50.	Hijo.	68.	Tiempo.
15.	Ridiculizar.	33.	Poder.	51.	Blog.	69.	Matón.
16.	Rumor.	34.	Venganza.	52.	Chat.	70.	Frecuente.
17.	Sufre.	35.	Acechar.	53.	Correo.	71.	Creciente.
18.	Suicidio.	36.	Escuela.	54.	Digital.		

El **corpus lingüístico de Ciberbullying** es sustentado en los libros: “Depredador Escolar Bully y Ciberbully, (Santander & Torres, 2013)” y “Acosados: Lo que los padres y maestros deben saber sobre el Bullying (Cobo & Tello, 2011)”.

3.2. Estudios de máquinas con aprendizaje supervisado.

Analizar y describir distintas máquinas con aprendizaje supervisado, considerar los siguientes algoritmos.

3.2.1. Árboles de decisión.

Es uno de los algoritmos más conocidos dentro de las distintas técnicas de aprendizaje supervisado; este se encarga de seleccionar el mejor atributo para la

raíz del árbol, divide el conjunto de ejemplos en conjuntos separados y agrega los nodos y ramas correspondientes al árbol, como se muestra en la Fig.2.

Los árboles de decisión aprenden de forma descendente, con un algoritmo conocido como inducción descendente de árboles de decisión (TDIDT), estos realizan una partición recursiva o aprendizaje de dividir y conquistar (Su, 2009).

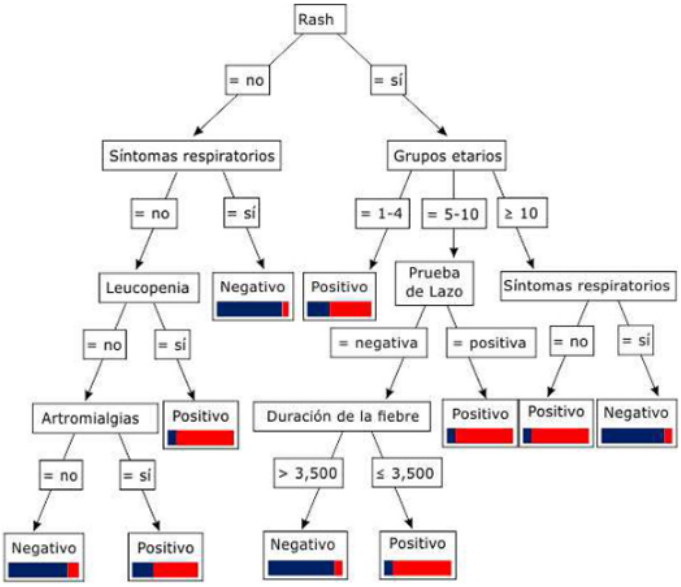


Figura 2. Modelo de árbol para síntomas y signos (Torres et al., 2016).

Según (Liberman, 2017), las desventajas de los árboles de decisión es que requieren de algoritmos capaces de determinar una elección óptima en cada nodo. Esto plantea que toma la decisión más óptima en cada paso, pero no tiene en cuenta el óptimo global. No obstante, elegir el mejor resultado en un paso dado no certifica que se dirija hacia una decisión óptima cuando llegue al último nodo del árbol, llamado nodo hoja (Liberman, 2017).

3.2.2. Métodos Naive Bayes.

Según (Gutiérrez Esparza, Margain Fuentes, Canul Reich, & Ramírez del Real, 2017), el algoritmo de Naïve Bayes está basado en el teorema de Bayes (1763) y en la premisa de independencia de los atributos dada una clase. La Fig. 3, describe un ejemplo del modelo Naive Bayes.

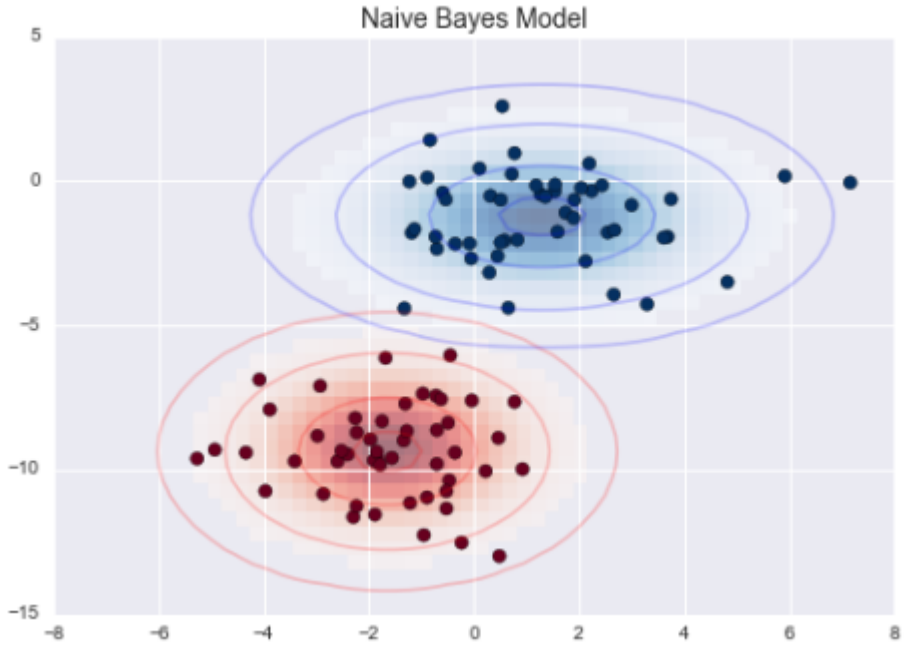


*Figura 3.*Ejemplo del modelo Naive Bayes (VanderPlas, 2016).

Asimismo, es uno de los métodos de aprendizaje supervisado más utilizados debido a que es posible adaptar para el análisis de emociones (Kaur & Singla, 2016; Tan, Cheng, Wang, & Xu, 2009).

En ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto

de datos no está en absoluto relacionada con la presencia de cualquier otra característica (Roman, 2018).

De igual manera, los puntos fuertes y débiles del algoritmo de Bayes son:

Fuertes:

- Un manera fácil y rápida de predecir clases, para problemas de clasificación binarios y multiclase.
- El desacoplamiento de las distribuciones de características condicionales de clase significa que cada distribución puede ser estimada independientemente como si tuviera una sola dimensión.

Débiles:

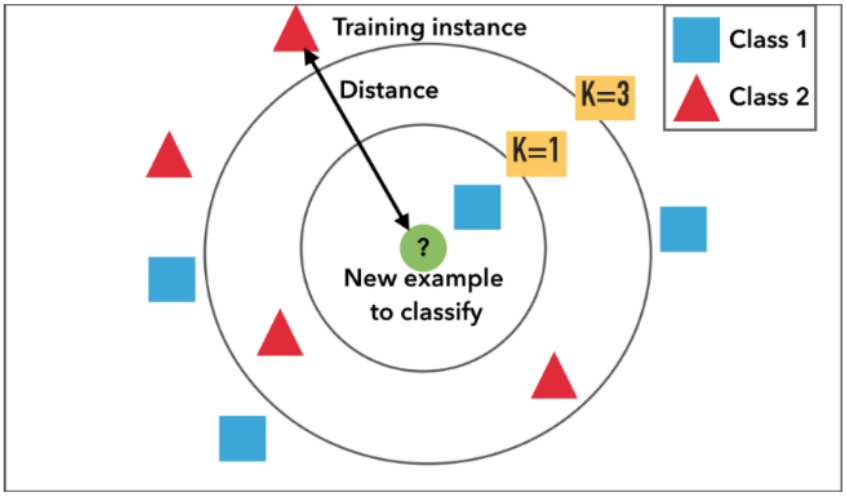
- Aunque son unos clasificadores bastante buenos, los algoritmos Naive Bayes son conocidos por ser pobres estimadores. Por ello, no se deben tomar muy en serio las probabilidades que se obtienen.
- La presunción de independencia Naive muy probablemente no reflejará cómo son los datos en el mundo real (Roman, 2018).

3.2.3. El K-ésimo vecino más cercano.

El algoritmo en este método, clasifica un objeto en una clase distinta cuando hay dos o más grupos de objetos de clase conocida. Se basa en el concepto de proximidad y no hace ninguna suposición acerca de la distribución en las clases (Miller & Miller, 2005).

Esta técnica de similitud asume que mientras más cerca estén los objetos en el espacio de medición, es más probable que pertenezcan a la misma categoría

o que sean similares con respecto a las variables en estudio, como se aprecia en la Fig. 4. En general, cuando el número de mediciones es dos o tres, la clasificación de los objetos se puede realizar mediante enfoques gráficos simples en dos o tres dimensiones, respectivamente (Vidueira Ferreira et al., 2015).



Ejemplo de clasificación k -NN. La muestra de prueba (círculo interior) debe clasificarse ya sea para la primera clase de cuadrados azules o para la segunda clase de triángulos rojos. Si k = 3 (círculo exterior) se asigna a la segunda clase porque hay 2 triángulos y solo 1 cuadrado dentro del círculo interior. Si, por ejemplo, k = 5 , se asigna a la primera clase (3 cuadrados vs. 2 triángulos fuera del círculo exterior).

Figura 4. Ejemplo K-nn Vecino más cercano (Bronshtein, 2017).

Algunas características de este algoritmo las explica de manera sencilla (Aler Mur, 2012). KNN: Clasifica nuevas instancias como la clase mayoritaria de entre los k vecinos más cercanos de entre los datos de entrenamiento; por otro lado, KNN es un algoritmo perezoso, esto se debe a que, durante el entrenamiento, sólo guarda las instancias, no construye ningún modelo (a diferencia de los árboles de decisión).

La clasificación se hace cuando llega la instancia de test, es no paramétrico

(no hace suposiciones sobre la distribución que siguen los datos, a diferencia de un modelo lineal) y por último el mejor modelo de los datos son los propios datos (Aler Mur, 2012).

3.2.4. Redes neuronales.

Las redes neuronales artificiales son modelos matemáticos que buscan replicar el comportamiento de las neuronas en la naturaleza, organizando su estructura de manera que emulen al cerebro. El objetivo de estas redes es lograr dar lugar a un sistema inteligente que logre realizar con éxito tareas complejas. Su funcionamiento no está orientado a ser similar al de un ordenador cuando procesa información, sino que persigue aproximarse a la inteligencia artificial. Así, las redes neuronales artificiales son capaces de crear patrones, reconocer información o resolver enigmas complejos (T Systems, 2017).

El primer modelo matemático de una neuronal artificial, creado con el fin de llevar a cabo tareas simples, fue presentado en el año 1943 en un trabajo conjunto entre el psiquiatra y neuroanatomista Warren McCulloch y el matemático Walter Pitts (Sancho Caparrini, 2018).

La Fig. 5, muestra un ejemplo de modelo neuronal con n entradas, que consta de:

- Un conjunto de entradas x1,…xn.
- Los pesos sinápticos w1,…wn, correspondientes a cada entrada.

- Una función de agregación, Σ.
- Una función de activación, f.
- Una salida, Y.

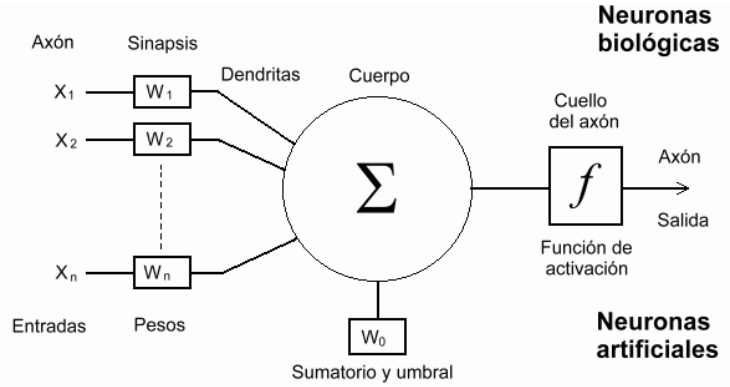


Figura 5 Modelo de Red Neuronal de McCulloch-Pitts (Sancho Caparrini, 2015).

3.2.5. Máquinas de Soporte Vectorial (MSV).

Para (Zhang, 2017), las Máquinas de Soporte Vectorial son una clase de algoritmos lineales que se pueden usar para la clasificación, regresión, estimación de densidad, detección de novedades, etc. En el caso más simple de clasificación de dos clases, las MSV encuentran un hiperplano.

La descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases (Bentacourt, 2005).

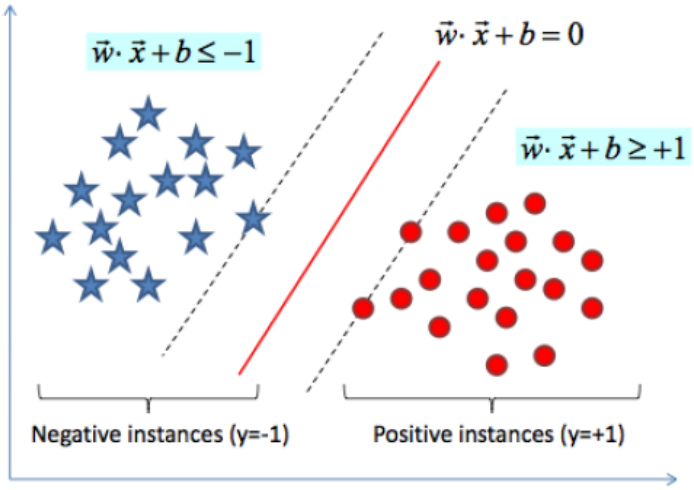


Figura 6. Ejemplos de aplicación de SVMs (Morales, González, & Hugo, 2017).

La idea principal de las MSV es encontrar el óptimo valor para esto separa el hiperplano entre las muestras positivas y negativas.(Ding, Yu, Qi, & Huang, 2014), como lo muestra la Fig. 6.

Las MSV puede ser más precisas en datos moderadamente desequilibrados. Sin embargo, las MSV son sensibles a una alta clasificación desequilibrada, ya que es propensa a generar un clasificador que tenga un fuerte sesgo de estimación hacia la clase mayoritaria y daría una mala precisión en el rendimiento de clasificación para la clase minoritaria (Wang & Xue, 2014).

3.2.6. Métodos de conceso (Random forest).

Para Fúrnkranz Random Forests es una técnica de aprendizaje en conjunto. El cual es un híbrido del algoritmo de empaquetamiento y del método de subespacio

aleatorio, y utiliza los árboles de decisión como el clasificador de base. Cada árbol se construye a partir de una muestra bootstrap del conjunto de datos originales.

Cada árbol se construye a partir de una muestra de arranque del conjunto de datos original. Un punto importante es que los árboles no están sujetos a podas después de la construcción, lo que les permite estar parcialmente atados a su propia muestra de los datos. Para diversificar aún más los clasificadores, en cada rama del árbol, la decisión de qué característica dividir está restringida a un subconjunto aleatorio de tamaño n, del conjunto de características completo (Fürnkranz, 2017).

Cada muestra del conjunto de datos de prueba está predicha por todos los árboles de decisión, y el resultado de la clasificación final se devuelve en función de los votos de estos árboles.

El conjunto de datos de entrenamiento original se formaliza como S = {(x i, y j), i = 1, 2, ..., N; j = 1, 2, ..., M}, donde x es una muestra e y es una variable característica de S. Es decir, el conjunto de datos de entrenamiento original contiene N muestras, y hay M variables de características en cada muestra (Chen et al., 2017).

La Fig. 7, presenta el proceso principal de la construcción del algoritmo de Random Forest.

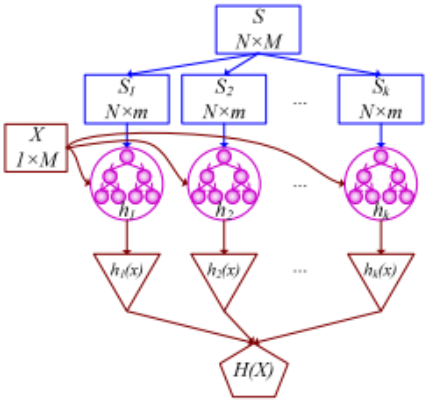


Figura 7. Ejemplo de construcción de Random forest (Chen et al., 2017).

3.2.7. Ada boosting.

AdaBoost es una contracción de “Adaptive Boosting”, en donde el término Adaptive hace alusión a su principal diferencia con su predecesor. El término boosting hace referencia a un tipo de algoritmos cuya finalidad es encontrar una hipótesis fuerte a partir de utilizar hipótesis simples y débiles (Breiman, Cutler, Cutler, & Stevens, 2001).

En términos de funcionalidad son iguales, ambos algoritmos buscan crear un clasificador fuerte cuya base sea la combinación lineal de clasificadores “débiles simples” h t (x). Sin embargo, AdaBoost propone entrenar una serie de clasificadores débiles de manera iterativa, de modo que cada nuevo clasificador o “weak learner” se enfoque en los datos que fueron erróneamente clasificados por su predecesor, de esta manera el algoritmo se adapta y logra obtener mejores resultados (William L. Hosch, 2009). La Fig.8, muestra un ejemplo gráfico de Ada Boost.

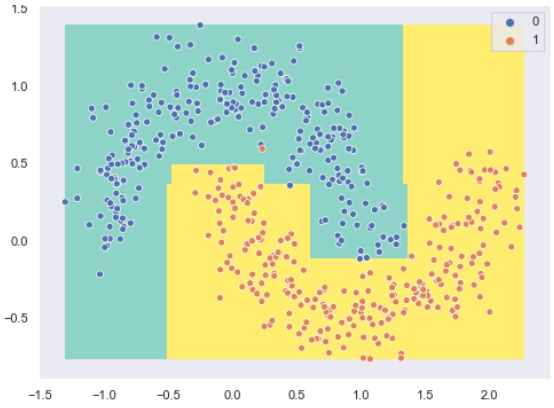


Figura 8. Ejemplo Adaboost (InteractiveChaos, 2017).

AdaBoost es uno de los primeros algoritmos que combinan múltiples "clasificadores débiles" en un solo "clasificador fuerte". Algunas características de Adaboost son:

- Los aprendices débiles en AdaBoost son árboles de decisión con una sola división, llamados troncos de decisión.
- AdaBoost funciona poniendo más peso en instancias difíciles de clasificar y menos en aquellas que ya se manejan bien.
- Los algoritmos de AdaBoost se pueden usar tanto para problemas de clasificación como de regresión (Desarda, 2019).

3.3. Diseñar el modelo de clasificación.

La Fig. 9, presenta el modelo de clasificación sobre el proceso que se lleva a cabo para la detección del vocabulario de Ciberbullying en Internet. El cual integra: Búsquedas inteligentes en la Web, herramienta ADVI para el pre procesamiento de

datos, detección de vocabulario con técnicas de aprendizaje de máquina y conocimiento.

Figura 9. Modelo de clasificación usado para la detección de Ciberbullying.

3.4. Evaluación del modelo de clasificación.

3.4.1. Búsquedas Inteligentes.

Las búsquedas inteligentes, se basan en un Crawler o rastreador Web, el cual se encarga de realizar exploraciones en la Web de manera automática. Su función consiste en descargar y añadir a un repositorio local las páginas visitadas, extrae los enlaces que contengan la página Web y los agrega a la lista de URLs, posteriormente analiza las páginas, y busca enlaces a sitios nuevos, repitiendo el

proceso hasta que se obtenga una muestra considerable de páginas Web para las pruebas.

3.4.2. Minería de datos (Data Mining).

La definición de la minería de texto según el autor (Jo, 2018), se conoce como el proceso de extraer el conocimiento implícito de los datos textuales. La clasificación, agrupación y asociación de textos, son las tareas típicas de la minería de textos, minería Web y la minería de datos de gran tamaño. El texto se define como los datos no estructurados que consisten en cadenas llamadas palabras.

3.4.3. Herramienta ADVI para el pre procesamiento de datos.

Un Dataset es un conjunto de datos utilizado para realizar las pruebas para las validaciones en la detección y predicción de vocabulario de Ciberbullying. La investigación actual, está basada en la investigación del Dr. (Castillo Zuñiga, I., Luna Rosas, F., Muñoz Arteaga, J., Lopez Veyna, 2016), la cual presenta una Arquitectura de software con el propósito de analizar grandes cantidades de información de la Web para apoyar la toma de decisiones de cualquier organización, aplicando métodos de Big Data Analitics y Web Semántica. Esta Arquitectura de Software consiste en obtener datos desde la web y finaliza con la detección de vocabulario (Conocimiento), aplicando diferentes técnicas demostrando con esto

que los porcentajes en la detección de vocabulario son muy altos, alcanzado un 95% de precisión. Los usos de las ontologías semánticas facilitan el análisis de la información, generando conjuntos de datos con diferentes perspectivas.

3.4.4. Detección de vocabulario con técnicas de Aprendizaje de Máquina.

Las técnicas usadas en el presente documento sobre los trabajos relacionados fueron las siguientes: K-ésimo vecino más cercano, métodos de conceso (Random Forest y Ada Boosting), Naive Bayes, Redes Neuronales, Máquinas de Soporte Vectorial, Arboles de Decisión. (Madala, Gangal, Krishna, Goyal, & Sureka, 2018; Mbaziira & Jones, 2016; Moohebat, Raj, Kareem, & Thorleuchter, 2015; Rodríguez Rama, 2018).

3.4.5. Técnicas, análisis y optimización de los datos.

Sobre los resultados que se obtienen en las pruebas, aplicando diferentes técnicas, se comprueba cuáles son las más adecuadas en el desarrollo de la investigación, y obtener un mayor porcentaje de aceptabilidad y precisión, teniendo la posibilidad de optimizar el tiempo de procesamiento en la recuperación del vocabulario y datos robustos para las pruebas.

3.4.6. Reporte de datos – predicciones.

Para generar el reporte de predicción, se realiza cuando se han ejecutado y unido las distintas técnicas de manera natural. Finalmente se prueba el dataset para validar los resultados a través de los algoritmos de Aprendizaje de Máquina en el lenguaje de programación R, el cual es uno de los lenguajes más usados en programación científica en el campo de la minería de datos, contando con diferentes bibliotecas o paquetes con funcionalidades de cálculo y graficación.

3.5 Comparar el modelo con otros modelos similares usados en el estado del arte.

Teniendo en cuenta el presente documento, podemos realizar una comparación de los resultados obtenidos de (Anggraini et al., 2018) el cual argumenta que su investigación trata sobre acoso cibernético en las redes sociales más usadas como lo es Twitter y sus posibles consecuencias. Para ello, se aplicó minería de texto a los tuits encontrando dos clases: el acoso y el no acoso. Se le aplico el método de Naive Bayes para evaluar su rendimiento utilizando para ello una matriz de confusión, arrojando un 94,75% de retirada y un 82,50% de precisión con +/-5,12% de desviación estándar, demostrando que es un modelo estable.

El trabajo desarrollado por (Álvarez-García, Barreiro-Collazo, Núñez, & Dobarro, 2015), demuestra que cuando se trata de niños menores hay que realizar

de base un proceso de resocialización. El objetivo de este trabajo es que con una herramienta se pueden realizar una serie de cuestionarios llamados CYBA, la cual nos proporcionará un estudio sobre las agresiones que se pueden realizar por medio de los diferentes aparatos tecnológicos que manejan estos menores, obteniendo consecuencias legales y psicológicas negativas. Debido a los avances tecnológicos, los jóvenes socializan de manera virtual, generándose agresiones más graves que las que se pueden generar cara a cara.

En Colombia podemos encontrar el artículo de (Rincón Rueda, Alberto Isaac; Ávila Díaz, 2014), el cual plantea, que debido al excesivo contacto de los jóvenes con la tecnología los distanciará más del mundo real. Su principal inquietud es saber por qué el acoso cibernético ha crecido en nuestro país y cómo podemos minimizar su impacto psicológico. Se pudo determinar que una de las causas es la permanencia en las redes sociales, convirtiéndose los jóvenes en un simple código de identificación, sumergiéndose en el consumismo que traen las tecnologías, las cuales no hacen distinción entre raza, sexo o condición humana. Esto ha permitido que las personas sean dominadas generando violencia contra los derechos humanos, por consiguiente, ha hecho que las autoridades creen normas para el buen uso de las tecnologías y hacer una especie de prevención sobre el acoso cibernético.

El artículo de (Raisi & Huang, 2018), manifiesta que el ciberacoso se está convirtiendo en una amenaza social para salud, dañando la vida de las personas. En este artículo se ilustra un modelo que trata sobre las frases claves identificadas por PVC. Tenemos varias formas de acoso cibernético como es la difusión de rumores y publicación de imágenes vergonzosas. El método consiste en máquinas de aprendizaje supervisado, el cual debe de ser capaz de encontrar patrones en los datos no etiquetados, el objetivo en este artículo es el de mejorar la detección automatizada del ciberacoso, como ejemplo de detección automatizada tenemos el de detectar secuencias de conversaciones.

El ciberacoso, de acuerdo con (Ortega Ruiz et al., 2016), en su escrito, lo menciona como un cara a cara a través de dispositivos digitales. El método que se usa en este documento es el de unas ecuaciones estructurales que evalúan la concurrencia y la relación entre bullying y Ciberbullying, prevaleciendo el primero sobre el segundo. Se puede realizar una intervención psicoeducativa y la psicología escolar para trabajar estos dos fenómenos.

Se plantea como un objetivo en el artículo de los autores (Herrera-López et al., 2017), la adaptación y la comprobación sobre las propiedades psicométricas de la escala del bullying EBIP-Q para Colombia, y se obtiene una muestra de 1931 adolescentes, con un 53% de mujeres, realizándose para este caso un análisis aplicado a los ítems, análisis factoriales confirmatorios y unos modelos de

ecuaciones estructurales. Se configura un segundo objetivo en contrar la prevalencia del Bullying y del Ciberbullying en Colombia y un tercer objetivo fue el de analizar el grado de relación y concurrencia de estos dos fenómenos. Se valoraron unos criterios de adecuación del vocabulario, claridad conceptual, coherencia y relevancia de cada ítem, realizándose una prueba de 60 estudiantes. Se manifiesta la necesidad de intervenir desde modelos ecológicos y globales la convivencia directa y la ciber-convivencia sin impedir acciones específicas de estos dos fenómenos.

3.6. Herramientas usadas en la analítica de Big Data.

Se describen cada una de las herramientas usadas en la solución del problema planteado, tanto de hardware como de software.

3.6.1 Hardware.

El computador con que se trabajó en las pruebas de la presente investigación y el uso de las diferentes técnicas para obtener resultados usando la minería de datos aplicado al texto, son los siguientes:

- **Marca:** Lenovo.
- **Modelo:** Z470.
- **Procesador:** Intel Core i7, 2,7 Ghz.
- **Ram:** 8 Giga Bytes.
- **Disco Duro:** Sata 500 GB y SSD 240 GB.

3.6.2 Software.

El equipo tiene las siguientes características en software:

- **Tipo de Sistema:** 64 Bits.
- **Sistema Operativo:** Windows 7 Profesional.
- **Programas:** Office 2016.
- **Lenguaje:** R Versión 3.6.0.
- **Lenguaje:** RStudio Version 1.2.1335.

Capítulo IV. Resultados y discusión.

4.1. Procedimiento general del ensayo.

En este proyecto de investigación se tomó como planteamiento importante el Ciberbullying, el cual es definido por los autores (Torres, Salas, & Torres, 2015), como: “El uso de las TIC como un medio para cometer acoso escolar entre compañeros”. El Ciberbulling se presenta en colegios de bachillerato, en especial en los grupos de grado avanzados, las más acosadas son las mujeres, los más acosadores son los hombres, este acoso se da más en colegios de estratos altos. Este tipo de acoso genera en los jóvenes ansiedad y depresión a corto y largo plazo, ocasionando bajo rendimiento académico y hasta la deserción de los estudiantes en los colegios. Debido al crecimiento exagerado de la tecnología y que los jóvenes son los principales consumidores, debemos de ponernos atentos y controlar el uso de los dispositivos para evitar desenlaces fatales.

Para la obtención del conjunto de datos de prueba (dataset), se implementó la arquitectura ADVI (Arquitectura para el análisis de datos orientados a Vocabularios en Internet), la cual rastrea los sitios Web de Ciberbullying, realiza una copia, y a través de un pre procesamiento de datos con técnicas de Web Semántica y Procesamiento de Lenguaje Natural, se obtiene el conjunto de datos que será usado en las pruebas de detección de Ciberbullying de esta investigación.

Se realiza la exploración de la información del dataset, en donde se establecen unas dimensiones para las pruebas que se le van a realizar usando pequeñas muestras de datos, generándo un sumario estadístico de las variables

constatando la integridad de la información. Los errores que se puedan presentar en los resultados como datos sucios, como pueden ser registros vacíos en el archivo de Excel, en datos numéricos encontrar alfanuméricos y datos que no están que aparezcan en las predicciones, etc.

El total de páginas web contenidas en el dataset son 1.064 registros, con 71 características usadas para la predicción, finalmente la variable categórica a predecir que contienen el “Si” o el “No”. Estas páginas web encontradas y descargadas por el crawler llamado ADVI (Castillo Zuñiga, I., Luna Rosas, F., Muñoz Arteaga, J., Lopez Veyna, 2016). La distribución de las muestras las encontramos de la siguiente manera, para la clase “Si” corresponde un valor de 649 y para la clase “No” da como resultado un valor de 415, para el entrenamiento de los datos se usó el 70% y para la prueba se usó el 30% de los datos del dataset, que son las páginas web que se almacenaron en este. Para el proceso de prueba se utilizó un valor constante que corresponde a la semilla, la cual nos garantiza en los datos su reproductividad.

4.2. Exploración y preparación de datos.

Para la realización de las pruebas se trabajó con un lenguaje de programación estadístico y gráfico llamado R, según (Ferrero, 2018), este software es usado en investigación científica cargando diferentes librerías o paquetes con funcionalidades de cálculo y graficación. Para fortalecer el ambiente, se usó RStudio como un entorno de pruebas y desarrollo.

Para realizar las pruebas se usó el 70% de los datos para el aprendizaje y el 30% para realizar el testing. La Fig.10 y11, presenta el vocabulario utilizado para las pruebas de Ciberbullying, y en la Fig. 12, presenta un resumen de la información por cada vocablo, ambas imágenes desplegadas en el lenguaje R.

Con la obtención de la matriz de confusión, la que nos permite comparar el número de aciertos y errores que nos generan las distintas clases, la cual nos arroja como resultado el porcentaje de acierto de cada modelo usado, estos valores se resumen en una tabla mostrando los resultados arrojados por cada técnica, donde se visualiza cual es la mejor técnica.

```
---------------Exploratory Data Analysis-------------------------------------####
Initial Exploratory Data Analysis

setwd("D:/INFORMACION/Documentos/Maestria/17.- Tesis/1.- Materias/Unidad 1/Entregas
      /Dataset/FilesLuis/DATASET")
datos<-read.csv("ciberbullying.csv",sep = ",",dec='.',header=T)
dimensiones
dim(datos)

 Vistazo general de como luce el dataset
 head(datos)
 Resumen de los datos
summary(datos)
```

Figura 10. Exploración de datos, fuente propia

```
> # Vistazo general de como luce el dataset
> head(datos)
  agresivo aislamiento amenaza ansiedad ataque autoestima ciberbullying
1       83           0       0        4      0         11             0
2        0           0       0        5      0          4           159
3        0           1       0        5      2          4             0
4        0           3       0        3      0          2             3
5        0          13       0        9      0         10             8
6        0           0       0        3      0          4            34
  depresion estres hostigar humillar insultos intimidar manipular ridiculizar
1         1      5        0        0        9         0         0            0
2         2      2        0        0        1         0         0            0
3         1      2        0        0        5         0         0            2
4         2      6        0        0        5         0         0            0
5         5      5        0       13        8         0         0            1
6         4      0        0        0        0         0         0            0
```

Figura 11. Pantallazo con los valores del Dataset generado en RStudio.

```
> # Resumen de los datos
> summary(datos)
    agresivo        aislamiento        amenaza          ansiedad
 Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
 1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   : 0.4558   Mean   : 0.1259   Mean   : 0.1664   Mean   : 0.3045
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   :91.0000   Max.   :13.0000   Max.   :21.0000   Max.   :51.0000
     ataque          autoestima      ciberbullying       depresion
 Min.   :0.00000   Min.   : 0.0000   Min.   :  0.000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:  0.000   1st Qu.:0.0000
 Median :0.00000   Median : 0.0000   Median :  1.000   Median :0.0000
 Mean   :0.09023   Mean   : 0.3863   Mean   :  3.614   Mean   :0.2519
 3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:  4.000   3rd Qu.:0.0000
 Max.   :6.00000   Max.   :44.0000   Max.   :159.000   Max.   :7.0000
```

Figura 12. Resumen de la información por cada vocablo. Fuente propia.

Para la preparación de los datos que van a ser estimados, se determinan el porcentaje de los datos que van a dedicarse al entrenamiento y los que van usarse para la prueba, se define la semilla inicial, la cual es el punto de partida donde se extrae de manera aleatoria la cantidad de datos que van a usarse para la prueba. Para generar los valores aleatorios se basó en una semilla, con el fin que siempre sean consistentes los números aleatorios. La Fig. 13, muestra el proceso mencionado.

```
##---------------Preparando Datos-------------------------------------####

NoRecords <- dim(datos)[1]
pctEntrenamiento <- 0.7
pctPrueba <- 0.3
semilla = 45

Cargamos dataset
setwd("D:/INFORMACION/Documentos/Maestria/17.- Tesis/1.- Materias/Unidad 1/Entregas/
      Dataset/FilesLuis/DATASET")datos<-read.csv("ciberbullying.csv",sep = ",",dec='.',header=T)
dim(datos)

con la funcion sample obtenemos una muestra aleatoria definida por una semilla que
permite un resultado repetible
set.seed(semilla)
muestra <- sample(NoRecords,pctEntrenamiento * NoRecords)
muestra

Generamos la tabla de aprendizaje
taprendizaje <- datos[muestra,]
dim(taprendizaje)
summary(taprendizaje)

Generamos la tabla de prueba con los datos que no corresponden a la muestra
ttesting <- datos[-muestra,]
dim(ttesting)
summary(ttesting)
```

Figura 13. Preparación de los datos a evaluar. Fuente Propia

4.3 Pruebas con los algoritmos de aprendizaje seleccionados.

4.3.1. Método Knn - Vecinos más cercanos.

El algoritmo Knn - vecino más cercano, consiste en clasificar un nuevo dato en su grupo correspondiente, dependiendo la cercanía del grupo calculando la distancia del nuevo respecto a los existentes, ordenando dichas distancias de menor a mayor, escogiendo el grupo al que pertenece, el cual corresponde al de mayor frecuencia y menor distancia.

La Fig. 14, muestra el código del algoritmo Knn – vecinos más cercanos, con los parámetros necesarios para poder probar el dataset. En el algoritmo se genera la función modelo a partir de la tabla de aprendizaje, posteriormente se

predicen los datos para el modelo de prueba y la matriz de confusión, como se muestra en la Fig.15.

```
##---------------Metodo KNN----------------------
#1.1 Metodo de los K vecinos mÃ¡s cercanos
suppressWarnings(suppressMessages(library(kknn)))

#Generamos la funcion modelo con la tabla de aprendizaje especificando
#la columna de la variable discriminante
modelo<-train.kknn(Correcta~.,data=taprendizaje,kmax=9)
modelo

#Generamos los valores predichos por el modelo en los datos de prueba
prediccion<-predict(modelo,ttesting[,-71])
prediccion

## Matriz de Confusion
MC<-table(ttesting[,71],prediccion)
MC

##Obtenemos el nivel de aciertos
acierto<-(sum(diag(MC)))/sum(MC)
acierto

#Obtenemos el error
error<-1-acierto
error
```

Figura 14. Código del algoritmo Knn – Vecinos más cercano. Fuente propia.

```
Type of response variable: nominal
Minimal misclassification: 0.3172043
Best kernel: optimal
Best k: 7
>
> #Generamos los valores predichos por el modelo en los datos de pureba
> prediccion<-predict(modelo,ttesting[,-71])
> prediccion
  [1] Si No Si Si Si Si No Si Si Si Si Si Si Si Si Si Si No No No Si Si Si Si Si Si Si Si Si No No Si Si Si Si Si Si No
 [39] Si Si No Si No Si Si No No Si Si Si Si No No No Si No No No Si Si No Si No Si No Si Si Si Si Si No No Si Si Si
 [77] No Si No Si Si Si No Si Si Si No Si No No No Si Si No No No No Si Si Si Si Si Si No No Si Si No Si Si Si No Si No
[115] No Si No No Si Si No Si Si No No Si Si Si Si No Si Si Si Si Si Si Si Si No Si No Si Si Si Si No Si Si Si Si No No
[153] No Si No Si No No Si No No Si Si No No No No No Si Si No Si Si No No Si No Si Si No No No Si Si No Si No Si No
[191] Si Si Si Si Si No No No Si Si No No No Si No Si Si Si No Si Si No No No Si Si Si Si Si No No Si Si No No No Si No
[229] No Si No No No Si No No No Si No Si Si No No Si Si No No No No Si No Si No Si Si No No No No No No No No No No
[267] No Si No Si No No No No No No No Si No No No No No No No No No No No No No No Si Si Si No No No Si No No No No No
[305] No No No No No No No No No No No No No No No No
Levels: no No Si
>
> ## Matriz de Confusion
> MC<-table(ttesting[,71],prediccion)
> MC
    prediccion
      no  No  Si
  no   0   1   0
  No   0  80  44
  Si   0  84 111
>
> ##Obtenemos el nivel de aciertos
> acierto<-(sum(diag(MC)))/sum(MC)
> acierto
[1] 0.596875
>
> #Obtenemos el error
> error<-1-acierto
> error
[1] 0.403125
> ##Fin K vecinos mas cercanos
> ##---------------Metodo C4.5----------------------------
```

Figura 15. Resultado del modelo del algoritmo Knn – Vecinos más cercano. Fuente propia

4.3.2. Método Naive Bayes.

Especificamos en este método la variable discriminante para que en el modelo, los datos de prueba se generen los valores predichos, al generarse la matriz de confusión se visualizaron los porcentajes de acierto y error respectivo. Al generarse la matriz de confusión, nos permite la comparación de los valores de acierto y error en diferentes clases, obteniendo como resultado los porcentajes de precisión de los modelos empleados. Con estos resultados calculados finalmente construimos una tabla en la cual podemos realizar la comparación de los resultados y apreciar de estas técnicas aplicadas cual es la que mejor resultados arroja, Fig. 16 y 17.

```
##---------------Metodo de Bayes-------------------------------
##1.2 Metodo de Bayes
suppressWarnings(suppressMessages(library(e1071)))

#Generamos el modelo Naive-Bayes, especificando la variable discriminante
modelo<-naiveBayes(Correcta~.,data=taprendizaje)
modelo

#Generamos los valores predichos por el modelo en los datos de prueba
prediccion<-predict(modelo,ttesting[,-71])
prediccion

## Matriz de Confusion
MC<-table(ttesting[,71],prediccion)
MC

##Obtenemos el nivel de aciertos
acierto<-(sum(diag(MC)))/sum(MC)
acierto

#Obtenemos el error
error<-1-acierto
error
##Fin Metodo Bayes
```

Figura 16. Código del algoritmo Naive Bayes, fuente propia

```
> #Generamos los valores predichos por el modelo en los datos de prueba
> prediccion<-predict(modelo,ttesting[,-71])
> prediccion
  [1] Si Si Si Si Si Si Si Si No Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si Si No Si Si
 [39] Si Si Si Si No Si Si Si Si Si Si Si Si No Si No Si Si No Si Si Si Si No Si No Si Si No Si No Si Si Si No Si No Si No
 [77] Si Si No No Si Si Si No Si No No No No Si Si No Si Si No No Si Si No No Si Si No Si No Si Si Si Si Si No No Si No
[115] Si No No Si No Si Si No No No No Si No No No No Si Si No Si No Si Si No No Si Si No Si No Si Si Si No No No No No
[153] Si Si No No No No Si No Si Si No No No No No No Si No No Si No No No No No No No No No Si Si No Si Si No No
[191] No No No No No No No No No No No No No No No No No No No No No No No No No Si No No No No No No No No No No
[229] No Si No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No
[267] No No No Si No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No No
[305] No No No No No No No No No No No No No No No
Levels: no No Si
```

Figura 17. Arroja el resultado del Modelo generado por Naive Bayes. Fuente propia

4.3.3. Método Random Forest.

En la *Fig. 18*, se construye y entrena el modelo con el algoritmo de Bosques Aleatorios, se valida su precisión con el conjunto de datos de prueba generándose la matriz de confusión con sus aciertos y errores, Fig.19.

```
##---------------Metodo de Random Forest------------------------------
#1.3 Metodo de Random Forest
suppressWarnings(suppressMessages(library(randomForest)))

#Generamos el modelo, especificando la variable discriminante
modelo<-randomForest(Correcta~.,data=taprendizaje,importance=TRUE)
modelo
#plot(modelo)

#Generamos los valores predichos por el modelo en los datos de prueba
prediccion<-predict(modelo,ttesting[,-71])
prediccion

## Matriz de Confusion
MC<-table(ttesting[,71],prediccion)
MC

# Porcentaje de buena clasificacion
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Random Forest
```

Figura 18 . Código del método Random Forest. Fuente propia

```
> ## Matriz de Confusion
> MC<-table(ttesting[,71],prediccion)
> MC
    prediccion
      no  No  Si
  no   0   1   0
  No   0 102  22
  Si   0 103  92
>
> # Porcentaje de buena clasificacion
> acierto<-(sum(diag(MC)))/sum(MC)
> acierto
[1] 0.60625
>
> error<-1-acierto
> error
[1] 0.39375
```

Figura 19. Muestra la Matriz de confusión y el porcentaje de aciertos del método de Random Forest. Fuente propia

4.3.4. Método Arboles de Decisión.

En la Fig. 20, se construye y entrena el modelo con el algoritmo Arboles de decisión, se valida su precisión con el conjunto de datos de prueba, se genera la matriz de confusión con sus aciertos y errores, Fig. 22.

```
##---------------Metodo de Arboles de Decision---------------
#1.4 Metodo de Arboles de Decision
suppressWarnings(suppressMessages(library(rpart)))
suppressWarnings(suppressMessages(library(rpart.plot)))

#Generamos el modelo, especificando la variable discriminante
modelo <- rpart(Correcta~.,data=taprendizaje)
modelo
plot(modelo)
text(modelo)
prp(modelo,extra=104,branch.type=2, box.col=c("pink", "palegreen3")[modelo$frame$yval])
plot(prp)
#Generamos los valores predichos por el modelo en los datos de prueba
prediccion <- predict(modelo, ttesting[,-71], type='class')
prediccion

## Matriz de Confusion
MC<-table(ttesting$Correcta,prediccion)
MC

# Porcentaje de buena clasificacion
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Arboles de decision
```

Figura 20. Se muestra el modelo Arboles de Decisión. Fuente propia

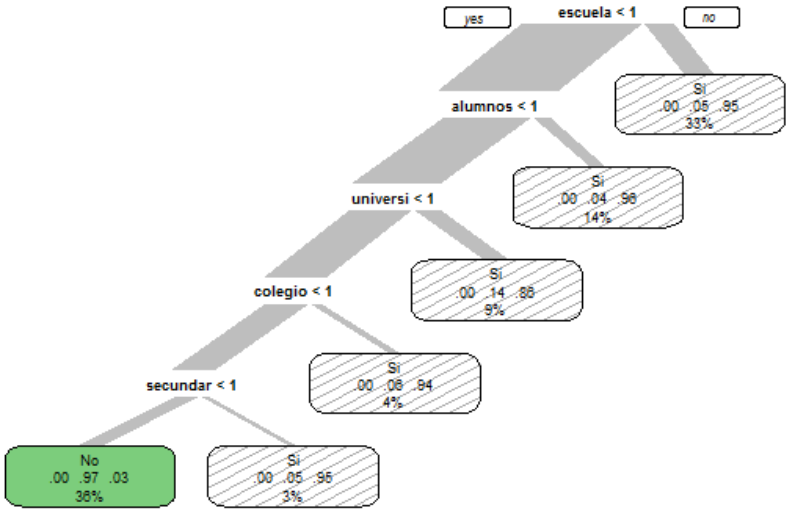


Figura 21. Arroja el árbol de decisión por vocablo evaluado. Fuente propia

```
> ## Matriz de Confusion
> MC<-table(ttesting$Correcta,prediccion)
> MC
    prediccion
      no  No  Si
  no   0   1   0
  No   0 105  19
  Si   0   6 189
>
> # Porcentaje de buena clasificacion
> acierto<-(sum(diag(MC)))/sum(MC)
> acierto
[1] 0.91875
>
> error<-1-acierto
> error
[1] 0.08125
```

Figura 22. Muestra el porcentaje de aciertos del método de árboles de decisión. Fuente propia.

4.3.5. Método Máquina de Soporte Vectorial.

En la Fig. 23, se construye y entrena el modelo con el algoritmo máquina de soporte vectorial, se valida su precisión con el conjunto de datos de prueba, se genera la matriz de confusión con sus aciertos y errores, como se observa en la Fig. 24.

```
##---------------Metodo Maquinas de Soporte Vectorial (SVM)--------------
#1.5 MÃ©todo Maquinas de Soporte Vectorial (SVM)
# El paquete 'kernlab' ofrece la funciÃ³n 'ksvm'.
suppressWarnings(suppressMessages(library(kernlab)))

#Generamos el modelo, especificando la variable discriminante
modelo <- ksvm(Correcta~.,
     data=taprendizaje,
     kernel="rbfdot", prob.model=TRUE, kpar=list(sigma=0.05),C=5,cross=3)
modelo


#Generamos los valores predichos por el modelo en los datos de pureba
prediccion <- predict(modelo, ttesting)
prediccion

## Matriz de Confusion
MC<-table(ttesting$Correcta,prediccion)
MC

# Porcentaje de buena clasificacion
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Maquinas de Soporte Vectorial (SVM)
```

Figura 23. Código del modelo gaussiano por ksvm. Fuente propia

```
> ## Matriz de Confusion
> MC<-table(ttesting$Correcta,prediccion)
> MC
    prediccion
      no  No  Si
  no   1   0   0
  No  60  64   0
  Si  21 174   0
>
> # Porcentaje de buena clasificacion
> acierto<-(sum(diag(MC)))/sum(MC)
> acierto
[1] 0.203125
>
> error<-1-acierto
> error
[1] 0.796875
```

Figura 24. Modelo del método ksvm. Fuente propia

4.3.6. Método Regresión Logística Lineal.

En la Fig. 25, se construye y entrena el modelo con el algoritmo Método regresión logística lineal, se valida su precisión con el conjunto de datos de prueba,

se genera la matriz de confusión con sus aciertos y errores, se muestra el código del modelo (Fig. 26).

```
##---------------Metodo de Arboles de Decision---------------
#1.4 Metodo de Arboles de Decision
suppressWarnings(suppressMessages(library(rpart)))
suppressWarnings(suppressMessages(library(rpart.plot)))

#Generamos el modelo, especificando la variable discriminante
modelo <- rpart(Correcta~.,data=taprendizaje)
modelo
plot(modelo)
text(modelo)
prp(modelo,extra=104,branch.type=2, box.col=c("pink", "palegreen3")[modelo$frame$yval])
plot(prp)
#Generamos los valores predichos por el modelo en los datos de prueba
prediccion <- predict(modelo, ttesting[,-71], type='class')
prediccion

## Matriz de Confusion
MC<-table(ttesting$Correcta,prediccion)
MC

# Porcentaje de buena clasificacion
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Arboles de decision
```

Figura 25. Código del Modelo de Regresión Logística Lineal. Fuente propia

```
> #Generamos el modelo, especificando la variable discriminante
> modelo <- glm(Correcta ~ .,
+     data=taprendizaje,
+     family=binomial(link="logit"))
> # summarize the fit
> summary(modelo)

Call:
glm(formula = Correcta ~ ., family = binomial(link = "logit"),
    data = taprendizaje)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4786  -0.5803   0.0004   0.4378   1.9809

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.900e+00  2.151e-01  -8.831  < 2e-16 ***
agresivo       8.996e-01  1.424e+00   0.632  0.52768
aislamiento    2.271e-01  5.712e-01   0.398  0.69098
```

Figura 26. Modelo generado por Regresión Logística Lineal. Fuente propia

4.3.7. Método Redes Neuronales.

En la Fig. 27, se construye y entrena el modelo con el algoritmo Método Redes neuronales, se valida su precisión con el conjunto de datos de prueba, se genera la matriz de confusión con sus aciertos y errores.

```
##---------------Metodo Redes Neuronales-----------------------------------
#1.7 Metodo Redes Neuronales
suppressWarnings(suppressMessages(library(nnet)))

#Generamos el modelo, especificando la variable discriminante
modelo <- nnet(as.factor(Correcta) ~ .,
               data=taprendizaje,
               size=10, skip=TRUE, MaxNWts=10000, trace=FALSE, maxit=100)
summary(modelo)


#Generamos los valores predichos por el modelo en los datos de prueba
prediccion <- predict(modelo, newdata=ttesting[,-71], type="class")

## Matriz de Confusion
MC<-table(ttesting[,71],prediccion)
MC

# Porcentaje de buena clasificaciÃ³n
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Redes Neuronales
```

Figura 27. Código del Método de Redes Neuronales. Fuente propia

4.3.8. Método Adaboost.

En la Fig. 28, se construye y entrena el modelo con el algoritmo Método Adaboost, se valida su precisión con el conjunto de datos de prueba, se genera la matriz de confusión con sus aciertos y errores.

```
##---------------Metodo Adaboost-----------------------------
#1.8 Metodo Adaboost
# El paquete 'adabag' ofrece la funcion 'boosting'.
suppressWarnings(suppressMessages(library(adabag)))

#Generamos el modelo, especificando la variable discriminante
modelo <- boosting(Correcta~., data=taprendizaje, boos=TRUE, mfinal=20)
modelo


#Generamos los valores predichos por el modelo en los datos de prueba
prediccion <- predict(modelo, ttesting)
prediccion

## Matriz de Confusion
MC<-table(ttesting$Correcta,prediccion$class)
MC

# Porcentaje de buena clasificaciÃ³n
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo Ada Boost
```

Figura 28. Código del método adaboost, fuente propia.

4.3.9. Método C 4.5.

En la Fig. 29, se construye y entrena el modelo con el algoritmo Método C 4.5, se valida su precisión con el conjunto de datos de prueba, se genera la matriz de confusión con sus aciertos y errores.

```
##---------------Me©todo C4.5-----------------------------
#1.9 Metodo C4.5
# Necesitamos Rweka para obtener el algoritmo J48(C4.5) en R
library(rJava)
library(RWeka)
library(caret)
modelo <- train(Correcta ~., method="J48", data=taprendizaje,
                tuneLength = 5,
                trControl = trainControl(
                  method="cv"))

#Generamos los valores predichos por el modelo en los datos de prueba
prediccion <- predict(modelo, ttesting)
prediccion

## Matriz de Confusion
MC<-table(ttesting$Correcta,prediccion)
MC

# Porcentaje de buena clasificacion
acierto<-(sum(diag(MC)))/sum(MC)
acierto

error<-1-acierto
error
## Fin del metodo C4.5
```

Figura 29. Código del método C 4.5, fuente propia

4.4. Resultados.

En la presente investigación se demuestra que el tener un conjunto de datos suficientemente preparado y depurado, se pueden implementar las distintas técnicas de machine learning para poder detectar los distintos vocablos referentes al Ciberbullying, lo cual se logra con el apoyo de la herramienta ADVI (Castillo Zuñiga, I., Luna Rosas, F., Muñoz Arteaga, J., Lopez Veyna, 2016).

Los métodos que presentaron valores suficientemente altos con porcentajes por encima al 85%, fueron el de Redes Neuronales y por encima del 90%, fueron Random Forest, Arboles de Decisión y Adaboost, esto genera un porcentaje aceptable en las distintas técnicas utilizadas; pero esto no oculta que hubo métodos que no presentaron valores que superaran el 60% como lo fueron el Método Knn-vecinos más cercanos y el método de Naive Bayes.

Después de realizar las distintas pruebas y consultar los distintos artículos y libros se puede argumentar que una de las grandes dificultades para la detección del vocabulario está dada en el conflicto para buscar y minar los datos que van a ser analizados y evaluados.

La técnica de machine learning que presento mayor porcentaje de acierto en la detección del vocabulario referente a Ciberbullyng en la presente investigación, fue la de Random Forest con un porcentaje de acierto superior al 93% y en su orden siguieron las técnicas de Adaboost con el 92,5% y Arboles de Decisión

con el 92,18%; en promedio estas 3 técnicas superan el 92% lo que garantiza una efectividad muy adecuada para la investigación.

4.5. Discusión.

Realizada la evaluación de las diferentes técnicas de Machine Learning al dataset, se describen los resultados en la Tabla 3, en donde se considera el método usado, el porcentaje de acierto y el porcentaje de error.

Tabla 3. Métodos usados en los algoritmos para calcular los porcentajes de acierto y error.

Método Usado	% Acierto	% de Error
K-vecinos más cercanos KNN	60,00	40,00
Naive Bayes	60,93	39,07
Random Forest	93,53	6,57
Arboles de Decisión	92,18	7,82
Máquinas de Soporte Vectorial (SVM)	73,43	26,57
Regresión Logística Lineal	73,43	26,57
Redes Neuronales	86,25	13,75
Adaboost	92,50	7,50
C4.5	92,18	7,82

De acuerdo a los resultados obtenidos en esta investigación, se obtiene 93.43% de precisión en la detección de vocabulario de Ciberbullying, el cual supera con un 11% de ventaja al trabajo presentado por el autor (Anggraini et al., 2018), que obtuvo un 82.50% de precisión.

4.5.1. Discusión de los objetivos.

De acuerdo al objetivo general que dice lo siguiente "*Aplicar técnicas de Machine Learning en el análisis de datos para predecir vocabulario de Ciberbullying, usando páginas de Internet descargadas por una aplicación para su posterior análisis*". La Tabla 3, muestra los algoritmos utilizados en las pruebas para la detección de Ciberbullying, comprobándose de manera indiscutible que se aplicaron en cada uno de los puntos planteados en el objetivo general, pudiéndose demostrar que las diferentes técnicas de aprendizaje supervisado, aplicadas al dataset son totalmente eficientes para poder detectar el vocabulario referente al Acoso escolar en Internet.

En cuanto a la discusión de los objetivos específicos se describen a continuación, los resultados obtenidos.

1. Utilizar distintas técnicas de aprendizaje de máquina para la detección de vocabulario de Ciberbullying.

En las pruebas con el dataset para la detección de Ciberbullying en sitios Web, se utilizaron nueve algoritmos de aprendizaje supervisado, Knn - vecino más cercano (60%), Naive Bayes (61%), Random Forest (94%), Arboles de decisión (92.18%), Máquinas de Soporte Vectorial (74%), Regresión Logística Lineal (74%), Redes Neuronales (87%), AdaBoost (93%), y C4.5 (92%), logrando así, probar

distintas técnicas para obtener una mejor precisión en la clasificación de las páginas Web de Ciberbullying.

2. Encontrar el algoritmo adecuado para clasificar de manera confiable la detección de vocabulario de Ciberbullying.

Con el algoritmo Bosques Aleatorios se alcanzó un porcentaje de precisión alto, del 94% de predicción en la clasificación de sitios Web de Internet sobre la temática tocante al Ciberbullying, seguido de los algoritmos AdaBoost, arboles de decisión y C4.5, con un porcentaje cercano al 93% de precisión.

4.5.2. Discusión de la hipótesis.

En esta investigación la hipótesis planteada dice los siguiente: "***Es posible predecir el vocabulario del Ciberbullying, clasificar sitios Web y obtener un valor agregado en Internet sobre las páginas que circulan en este, a través de la aplicación de técnicas de Análisis de Big Data, Procesamiento de Lenguaje Natural, Web Semántica y Aprendizaje de Máquina (Aprendizaje Supervisado)***", sobre este planteamiento se puede afirmar que el argumento que plantea la hipótesis es posible, debido a que se lograron obtener resultados altos en la detección del vocabulario de Ciberbullying en Internet, con una precisión del 94% al utilizar Bosques Aleatorios. Cabe señalar que la investigación sólo aplica para las palabras en el idioma español.

Capítulo V. Conclusiones.

5.1. Conclusiones generales.

En la presente investigación, se apunta a una serie de elementos para poder realizar una serie de análisis usando como tema principal el Ciberbullying, el cual es una forma de acoso cibernético que se da en preadolescentes y adolescentes, provocados por los avances tecnológicos en forma de contacto electrónico. Esto sucede en las redes sociales, donde las personas no pueden defenderse fácilmente, generando síntomas depresivos y de ansiedad, es afectada su autoestima, bajo rendimiento académico, se ven disminuidas sus relaciones interpersonales, frustración e indefensión, ausentismo escolar, el consumo de sustancias psicoactivas, conductas delictivas, entre otras. Para su detección, se aplicaron las técnicas de machine learning seleccionadas para la clasificación y predicción del vocabulario del Ciberbullying, las cuales son: Máquinas de Soporte Vectorial (SVM), Naive Bayes, Adaboost, Random Forest, , Regresión Lineal, C 4.5, Redes Neuronales, K-vecinos más cercanos Knn y Arboles de Decisión.

Los métodos que presentaron una menor efectividad o acierto respecto a la detección de vocablos o términos del Ciberbullying fueron: K-vecinos más cercanos Knn y Naive Bayes, con un porcentaje de acierto de aproximadamente del 60%, y los que presentaron el más alto acierto, fue el de Random Forest con un 93,43 %, siguiéndole en su orden los algoritmos Arboles de Decisión, Adaboost y C4,5, los cuales estuvieron entre el 92% y 93% de acierto.

5.2. Ventajas de la investigación.

Una de las ventajas que observamos en esta investigación, fue la variedad de algoritmos usados y la integración de los mismos, permitiendo un análisis profundo sobre la detección de vocabulario concerniente al Ciberbullying. En donde se obtiene como conocimiento, que la mejor técnica de aprendizaje supervisado, fue Random Forest, con un porcentaje de precisión del (93,43%), con una pequeña ventaja sobre otros tres algoritmos que fueron Arboles de Decisión (92,18%), Adaboost (92,50%) y el C 4.5 (92,18%).

5.3. Trabajos futuros.

Como trabajo futuro y de acuerdo a los resultados obtenidos, se pretende llevar a cabo un estudio con un conjunto de datos (dataset), que cuente con características para poder determinar en dónde está ocurriendo este tipo de acoso cibernético, quienes lo están generando, las edades de los muchachos, que medio tecnológico están usando, entre otros, con el propósito de poder reglamentar normas y realizar variedad de campañas de prevención y que se puedan prevenir este tipo agresiones.

Referencias.

Aler Mur, R. (2012). CLASIFICADORES KNN-I. *Universidad Carlos III*.

Álvarez-garcía, D., Barreiro-collazo, A., Nuñez, J. C., & Dobarro, A. (2015). Español Inglés. *The European Journal of Psychology Applied to Legal Contex*, 1–17.

Álvarez-García∗, D., Barreiro-Collazo, A., Núñez, J. C., & Dobarro, A. (2015). The European Journal of Psychology Applied to Legal Context. *The European Journal of Psychology Applied to Legal Contex*, *7*(1), 41–49. https://doi.org/10.1016/j.ejpal.2014.11.003

Anggraini, I. Y., Sucipto, S., & Indriati, R. (2018). Cyberbullying Detection Modelling at Twitter Social Networking. *JUITA : Jurnal Informatika*, *6*(2), 113. https://doi.org/10.30595/juita.v6i2.3350

Bentacourt, G. (2005). Las Máquinas de Soporte Vectorial (SVMs). *Scientia et Technica*, *1*(27), 67–72.

Breiman, L., Cutler, A., Cutler, D. R., & Stevens, J. R. (2001). Random forests. *Ensemble Machine Learning: Methods and Applications*, 5–32. https://doi.org/10.1007/9781441993267_5

Bronshtein, A. (2017). Una introducción rápida al algoritmo de los vecinos más cercanos.

Castillo Zuñiga, I., Luna Rosas, F., Muñoz Arteaga, J., Lopez Veyna, J. (2016). *Architecture (ADVI) for the detection of cyberbullying vocabulary in internet combining techniques of big data analytics and semantic. 3*, 12. https://doi.org/http://dx.doi.org/10.6036/NT8032

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., & Li, K. (2017). A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment. *IEEE Transactions on Parallel and Distributed Systems*, *28*(4), 919–933. https://doi.org/10.1109/TPDS.2016.2603511

Cobo, P., & Tello, R. (2011). *Acosados. Lo que los padres y maestros deben saber sobre el bullying*.

Corcoran, L., Guckin, C., & Prentice, G. (2015). Cyberbullying or Cyber Aggression?: A Review of Existing Definitions of Cyber-Based Peer-to-Peer Aggression. *Societies*, *5*(2), 245–255. https://doi.org/10.3390/soc5020245

Desarda, A. (2019). Entendiendo AdaBoost - Hacia la ciencia de datos.

Ding, S., Yu, J., Qi, B., & Huang, H. (2014). An overview on twin support vector machines. *Artificial Intelligence Review*, *42*(2), 245–252. https://doi.org/10.1007/s10462-012-9336-0

Erl, T., Khattak, W., & Buhler, P. (2014). *Big Data Fundamentals*.

Ferrero, R. (2018). Qué es R Software | Máxima Formación.

Fürnkranz, J. (2017). Encyclopedia of Machine Learning and Data Mining (2nd edition). In Springer (Ed.), *Reference Reviews* (2nd ed., Vol. 32). https://doi.org/10.1108/rr-05-2018-0084

González-Garcia, A. (2016). ARTÍCULO Factores de riesgo en el ciberacoso: revisión sistemática a partir del modelo del triple riesgo delictivo (TRD). *Ciberdelincuencia y Cibervictimización » de La Información*, *22*, 73–92.

Gutiérrez Esparza, G., Margain Fuentes, L., Canul Reich, J., & Ramírez del Real, T. A. (2017). Un modelo basado en el Clasificador Naïve Bayes para la evaluación del desempeño docente. *RIED. Revista Iberoamericana de Educación a Distancia*, *20*(2), 293. https://doi.org/10.5944/ried.20.2.17717

Herrera-López, M., Romera, E., & Ortega-Ruiz, R. (2017). Bullying y cyberbullying en Colombia; coocurrencia en adolescentes escolarizados. *Revista Latinoamericana de Psicologia*, *49*(3), 163–172. https://doi.org/10.1016/j.rlp.2016.08.001

InteractiveChaos. (2017). Ejemplo con AdaBoostClassifier | Interactive Chaos.

Jo, T. (2018). *(Studies in Big Data) Taeho Jo-Text Mining_ Concepts, Implementation, and Big Data Challenge-Springer (2018).pdf*.

Kaur, G., & Singla, A. (2016). *Sentimental Analysis of Flipkart reviews using Naïve Bayes and Decision Tree algorithm*. *5*(1), 148–153. Retrieved from http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-1-148-153.pdf

Lantz, B. (2013). *Machine Learning With R :Read this*. https://doi.org/10.1017/CBO9781107415324.004

Liberman, N. (2017). *Árboles de decisión y bosques aleatorios: hacia la ciencia de datos*. Retrieved from https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

Madala, D. S. V., Gangal, A., Krishna, S., Goyal, A., & Sureka, A. (2018). An

empirical analysis of machine learning models for automated essay grading. *PeerJ Preprints*, *6*, e3518v1. https://doi.org/10.7287/peerj.preprints.3518v1

Mahmood, Z. (2016). Data Science and Big Data Computing. In *Data Science and Big Data Computing*. https://doi.org/10.1007/978-3-319-31861-5

Martínez-Gamboa, R. (2016). Big Data en humanidades digitales: de la escritura digital a la "lectura distante." *Revista Chilena de Literatura*, (94), 39–58. Retrieved from http://www.revistaliteratura.uchile.cl/index.php/RCL/article/viewFile/44969/47051

Mbaziira, A., & Jones, J. (2016). *A Text-based Deception Detection Model for Cybercrime*.

Miller, J. N. (James N. ., & Miller, J. C. (Jane C. (2005). *Statistics and chemometrics for analytical chemistry*. Pearson/Prentice Hall.

Moohebat, M., Raj, R. G., Kareem, S. B. A., & Thorleuchter, D. (2015). Identifying ISI-indexed articles by their lexical usage: A text analysis approach. *Journal of the Association for Information Science and Technology*, *66*(3), 501–511. https://doi.org/10.1002/asi.23194

Morales, E., González, J., & Hugo, E. (2017). *Máquinas de Soporte Vectorial*. *2017*, 1–83.

Ortega Ruiz, R., Rey, R. Del, & Casas, J. A. (2016). Assessing bullying and cyberbullying: Spanish validation of EBIPQ and ECIPQ. *Psicología Educativa*, *22*, 71–79. https://doi.org/10.1016/j.pse.2016.01.004

Raisi, E., & Huang, B. (2018). Weakly supervised cyberbullying detection with participant‐vocabulary consistency. *Social Network Analysis and Mining*, *8*(1), 1–17. https://doi.org/10.1007/s13278-018-0517-y

Requena-Mesa, A. (2018). Big Data: La evolución de los datos. Retrieved from OpenWebinars.net website: https://openwebinars.net/blog/big-data-la-evolucion-de-los-datos/

Rincón Rueda, Alberto Isaac; Ávila Díaz, W. D. (2014). Ciberbullying En Colombia. *Revista Científica "General José María Córdova," 12*, 149–164.

Rodríguez Rama, J. M. (2018). *APLICACIÓN DE TÉCNICAS DE MACHINE*

LEARNING A LA DETECCIÓN DE ATAQUES. Retrieved from http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81126/11/jmrodriguez85TFM0618memoria.pdf

Roman, V. (2018). Algoritmos Naive Bayes: Fudamentos e Implementación.

Sancho Caparrini, F. (2015). *Introducción al Aprendizaje Automático*. 1–12.

Sancho Caparrini, F. (2018). Redes Neuronales: una visión superficial - Fernando Sancho Caparrini.

Santander, A. C., & Torres, J. V. (2013). *Depredador Escolar Bully y Ciberbully*.

Su, F. P. (2009). Patent Priority Network : Linking Patent Portfolio. *Journal of the American Society for Information Science*, *60*(1999), 2353–2361. https://doi.org/10.1002/asi

T Systems, . (2017). ¿Qué son las redes neuronales y cómo funcionan?

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). *Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis*. https://doi.org/10.1007/978-3-642-00958-7_31

TORRES, D., SALAS, L., & TORRES, J. (2015). Cyberbullying: El Acoso Escolar En La Era Virtual! *Instituto Alberto Merani*, 1–68. Retrieved from http://www.institutomerani.edu.co/publicaciones/tesis/2015/cyberbullying_acoso_escolar_era virtual.pdf

Torres, J. A., Meneses, L. O., Sokol, N., Sardiñas, R. B., Díaz, D. M., Sansón, R. B., & Arce, M. E. S. (2016). Técnica Árboles de decisión aplicada al método clínico en el diagnóstico del dengue. *Rev Cubana Pediatrica*, *88*.

Vandebosch, H. (2014). Minding Minors Wandering the Web: Regulating Online Child Safety. *Information Technology and Law Series*, *24*, 245–262. https://doi.org/10.1007/978-94-6265-005-3

Vandebosch, H., @bradleyboehmke, Abbass, H. A., Sarker, R. A., Newton, C. S., Aggarwal, C. C., … Elbeqqali, O. (2016). Combatting Cybercrime and Cyberterrorism: Challenges, Trends and Priorities. In S. Goel (Ed.), *Data Mining*. https://doi.org/10.1007/978-3-319-38930-1

VanderPlas, J. (2016). En profundidad: Clasificación ingenua de Bayes | Manual de ciencia de datos de Python.

Vidueira Ferreira, J. E., da Costa, C. H. S., de Miranda, R. M., de Figueiredo, A. F., Ferreira, J. E. V., da Costa, C. H. S., … de Figueiredo, A. F. (2015). The use of the k nearest neighbor method to classify the representative elements. *Educación Química*, *26*(3), 195–201. https://doi.org/10.1016/j.eq.2015.05.004

Wang, Z., & Xue, X. (2014). Support vector machines applications. *Support Vector Machines Applications*, *9783319023*, 1–302. https://doi.org/10.1007/978-3-319-02300-7

William L. Hosch. (2009). Clasificadores Débiles - AdaBoost. *Britannica Articles*, 21–31. Retrieved from https://www.britannica.com/technology/machine-learning#Article-History

Zhang, X. (2017). Support Vector Machines. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1214–1220). https://doi.org/10.1007/978-1-4899-7687-1_810