UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA

DETECCIÓN AUTOMÁTICA DE CIBERACOSO EN REDES SOCIALES

POR

LAURA PATRICIA DEL BOSQUE VEGA

COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE
DOCTOR EN INGENIERÍA CON ORIENTACIÓN
EN TECNOLOGÍAS DE LA INFORMACIÓN

MAYO, 2017

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE INGENIERÍA MECÁNICA Y ELÉCTRICA
SUBDIRECCIÓN DE ESTUDIOS DE POSGRADO

TESIS

DETECCIÓN AUTOMÁTICA DE CIBERACOSO EN REDES SOCIALES

POR

LAURA PATRICIA DEL BOSQUE VEGA

COMO REQUISITO PARCIAL PARA OBTENER EL GRADO DE
DOCTOR EN INGENIERÍA CON ORIENTACIÓN
EN TECNOLOGÍAS DE LA INFORMACIÓN

MAYO, 2017

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección de Estudios de Posgrado

Los miembros del Comité de Tesis recomendamos que la Tesis «Detección automática de ciberacoso en redes sociales», realizada por el alumno Laura Patricia Del Bosque Vega, con número de matrícula 1016569, sea aceptada para su defensa como requisito parcial para obtener el grado de Doctorado en Ingeniería con Orientación en Tecnologías de la Información.

El Comité de Tesis

Dra. Sara Elena Garza Villarreal
Asesor

Dr. Luis Martin Torres Trevino
Revisor

Dra. Lorena Beatriz Martínez Elizalde
Revisor

Dr. Héctor Gibrán Ceballos Cancino
Revisor

Dra. Valeria Paola González Dueñes
Revisor

Vo. Bo.

Dr. Simón Martínez Martínez
Subdirección de Estudios de Posgrado

San Nicolás de los Garza, Nuevo León, mayo 2017

Gracias Dios y a mi santa Madre María,
por estar conmigo....SIEMPRE.

A mi espOSO, Mario... GRACIAS por tu apoyo...we did it!

A mi familia de origen,
mis papás y hermanos y familia añadida, no lo hubiera logrado sin su ayuda.

A mi Vilos,
que desde el cielo me ven triunfar.

A mi familia del Encuentro de Novios,
gracias por la tolerancia!

A mi familia de FIME,
ellos saben quienes son.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Estoy mas que agradecida al comité de esta tesis, por su invaluable apoyo, eternamente agradecida a todos los doctores del programa de este doctorado, cada uno de los doctores me apoyaron de una manera muy especial, no tengo con que pagarles.

Mi reconocimiento y gratitud a mi alma matter, la Universidad Autónoma de Nuevo León y a la Facultad de Ingeniería Mecánica y Eléctrica por las facilidades otorgadas en las becas para el inicio y continuación de este estudio. Al igual de agradecida con la coordinación Nacional de Becas de Educación Superior por la beca otorgada por un año (2014-2015), gracias a ella tengo mi Macbook Pro la cual fue una herramienta fundamental para la realización de esta tesis.

A mi familia de FIME, en especial al Ing. Esteban Baez Villarreal que me permitió la oportunidad de empezar estos estudios en la facultad, siendo el director de la misma, al Dr. Jaime Arturo Castillo Elizondo por seguir apoyandome como director de la FIME, con las autorizaciones para las becas en mis estudios de posgrado, a mi querida Ing. Maria Elena Guerra, no tengo palabras para agradecerle el haberme apoyado con el tiempo de las clases, con mi horario, en fin con TODO. Al igual a quien puedo decir que es mi amigo, Ing. Jesus Adolfo Melendez Guevara «Mele», gracias por ayudarme al empezar y ahora a terminar, gracias por apoyarme en lo que estuvo en tus manos.

Y un párrafo para ellos solos a los mejores compañeros que alguien puede tener: Ing. Raquel Martinez «Rachel», Ing. Jose Luis Torres «JL», en especial a ustedes por darme ese gran impulso que requería, Ing. Karla Porras, por ayudarme a ver todo el panorama, Ing. Arturo Del Angel, por ayudarnos tanto, por no dejarnos dormirnos en nuestros laureles y por empujarnos solamente para adelante. En verdad Dios se los pague, ustedes mejor que

nadie saben por lo que pasé y por lo que deje de ser y hacer para que esto ocurriera.

A mi Familia, en especial a mi marido, Mario, gracias por apoyarme, gracias por no dejarme caer, gracias por aguantar mis desvelos, mis desmañanadas, mi cambio de humor, las decisiones, la casa boca arriba, las comidas fuera, las no comidas, en fin....you're the best hubby ever. Lo bueno de esto es que tengo toda una vida para agradecerte por tu apoyo incondicional.

Mis papás, mis hermanos, mi mami por ayudarme con la casa, cuando lo requería, Adriana por leer mi tesis y corregir lo que se tenia que corregir. Mi papá por estar ahi y decirme que contaba contigo. Coco, por no molestar. Magdely, por dejar que coco no molestara :) .

Y por ultimo pero no por ello menos importantes, a todos mis amigos y compadres que apoyaron y que sufrieron y que estudiaron junto conmigo: Compadres: Ada y Omar Hdz, por ayudarme con mis primeros programas; Adriana y Mauro Villarreal, Adri por ayudarme con la redacción de la tesis y escucharme y Mauro mil gracias literal por todo :). Lety y Diego, por ayudarme a distraerme con Guanajuato. Y con todos mis amigos del Encuentro de novios Monterrey que toleraron mis: no puedo, estoy dormida, dentro de 3 años platicamos, pero sobre todo GRACIAS por sus ORACIONES.

Resumen

Laura Patricia Del Bosque Vega.

Candidato para obtener el grado de Doctorado en Ingeniería con Orientación en Tecnologías de la Información.

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Título del estudio: Detección automática de ciberacoso en redes sociales

Número de páginas: 113

Objetivos y método de estudio: El objetivo general de esta investigación es el de contribuir al desarrollo de un enfoque que permita avanzar en la detección del ciberacoso de manera automática en una red social, utilizando técnicas de aprendizaje computacional, análisis de sentimiento y minería de datos, herramientas que forman parte de las tecnologías de información. De manera particular, para desarrollar este enfoque, se realiza una búsqueda de los comentarios destacados como agresivos. Además, se identifican los involucrados dentro de un caso de ciberacoso, así como la frecuencia con la que se envían los comentarios agresivos, siendo estos los componentes que se consideran para lograr la detección de ciberacoso en una red social.

Contribuciones y conlusiones: La contribucion principal es una metodología que favorece en la detección de casos de ciberacoso en una red social. Este proceso de búsqueda, comienza con la recopilación de comentarios y la asignación automática de un nivel de

agresividad a estos comentarios.

Este nivel de agresividad nos ayuda a poder identificar los componentes que se consideran en un caso de ciberacoso, la frecuencia del envío de mensajes de texto considerados agresivos y los involucrados en este envío de mensajes. Al contar con estos datos se puede lograr conseguir detectar casos de ciberacoso en una red social.

Firma del asesor: _____

Dra. Sara Elena Garza Villarreal

Capítulo 1

Introducción

Mi única esperanza es que nunca perdamos de vista una única cosa: Que todo empieza con una razón.

Walt Disney

La internet ha marcado una diferencia desde que surgió, cambiando la manera en la que las personas nos comunicamos, hablamos, opinamos e inclusive buscamos información (Castells, 2016). Esta diferencia ha sido positiva ya que une personas que se encuentran en continentes diferentes, reúne comunidades las cuales manejan gustos en común y no tienen que encontrarse en el mismo lugar para verse cara a cara (Kiesler, 2014). También en la internet se forman redes sociales de las cuales se puede obtener información (Kleinberg, 1999). Estos lazos se han hecho estrechos gracias a las tecnologías de información que apoyan a la comunicación que cambia drásticamente al poder hablar desde tu coche, desde un dispositivo electrónico y no solo hablar sino también ver a tus seres queridos (Wellman, 1999).

La manera en la que se comunican las personas hoy en día, sigue en constante evolución (Sticca y Perren, 2012). Aunque la tecnología proporciona muchos beneficios, también cuenta con efectos negativos, los cuales pueden dañar a las personas pero en especial a los jóvenes (Bravo y Rasco, 2013). Los jóvenes entre 13 y 19 años son los que utilizan con frecuencia el correo electrónico, mensajes de texto, chats, celulares inteligentes

(*Smartphones*), cámaras web y sitios web y estos pueden ser utilizados para molestar de manera grave a otras personas (Campbell, 2005). El equilibrio que existe con la tecnología moderna, entre los riesgos y las oportunidades, que puede presentárseles, se manifiesta claramente en un problema social que va en crecimiento, conocido como ciberacoso o *cyberbullying* (Walrave y Heirman, 2011). La presente investigación trata de la detección automática de ciberacoso en redes sociales.

1.1 Antecedentes

El acoso o "bullying" tradicional no es nada nuevo; ésto ha afectado a muchas generaciones (Tarapdar y Kellett, 2012); el acoso es considerado como un encuentro cara a cara entre niños y adolescentes en los patios de las escuelas, pero hoy en día este problema ha encontrado su camino hacia el ciberespacio (Dadvar y de Jong, 2012).

El ciberacoso o "cyberbullying" es definido como un acto agresivo, intencional realizado por un grupo o un individuo, utilizando vías electrónicas para contactar (por ejemplo, correo electrónico, chats, redes sociales), de manera consecutiva a una víctima que no puede defenderse fácilmente por ella misma (Espelage y Swearer, 2003). El ciberacoso es realizado a través de medios electrónicos de comunicación como: Messenger, Facebook, Twitter, Youtube, etc. además comparado con el acoso tradicional el ciberacoso es más agresivo (Sticca y Perren, 2012).

En un estudio realizado en México, el 20 % de estudiantes de preparatoria de una muestra de 1066 estudiantes fueron víctimas de insultos, humillaciones y acoso sexual por medios tecnológicos de comunicación (López Lucio, 2009); una cifra similar (cercano al 20 %) de estudiantes de secundaria de un colegio privado han sido acosados electrónicamente (Mendoza, 2012).

El Centro de Investigación de Ciberacoso en Estados Unidos ha encontrado las siguientes estadísticas (Webster, 2013)[2013]:

- Más del 80 % de adolescentes utiliza el celular con acceso a internet con regularidad, haciendo éste la tecnología más popular para cometer ciberacoso.

- Cerca de la mitad de los adolescentes han experimentado algún tipo de ciberacoso y 10 del 20 % lo vive regularmente.
- Las víctimas de ciberacoso tienen baja autoestima y consideran el suicidio.
- Los casos de suicidio en jóvenes y el ciberacoso están vinculados de manera importante.

En Londres, se han realizado varios estudios para determinar si el ciberacoso es una extensión del acoso tradicional fuera de la escuela y esta investigación dio como resultado que los casos de acoso presentados dentro de las instalaciones continuaban un 11 % afuera de ellas (Smith y Collage, 2006).

El ciberacoso ha sido asociado a experiencias negativas; diferentes estudios han demostrado que las víctimas de ciberacoso han reportado bajo aprovechamiento académico, las relaciones con la familia son de baja calidad, existen dificultades para relacionarse socialmente y desórdenes afectivos (Machmutow *et al.*, 2012; Tokunaga, 2010).

La diferencia entre el acoso tradicional y el ciberacoso se debe a la manera de contactar a la víctima, éste último de manera electrónica (Sticca y Perren, 2012); la diferencia conlleva a aspectos específicos de ciberacoso que se deriva al uso de medios electrónicos: un gran potencial de alcanzar una gran audiencia (publicidad), un gran potencial de no conocer la identidad del agresor (anonimato), poco nivel de retroalimentación directa con el agresor y la víctima; puede convertirse en una "bola de nieve" que puede llegar a salirse de control debido a la tecnología (Slonje *et al.*, 2012) que tiene bajos niveles de supervisión (Patchin y Hinduja, 2006), especialmente en las comunidades de las redes sociales (Dinakar *et al.*, 2012).

El rápido crecimiento de las redes sociales está alentando las actividades de ciberacoso (Nahar *et al.*, 2012); las implicaciones del ciberacoso se convierten en serias (intentos de suicidios) cuando las víctimas no pueden enfrentar la tensión emocional del abuso, el maltrato, la humillación y los mensajes agresivos (Campbell, 2005).

1.2 Motivación y Justificación

Como se ha mencionado, el *bullying* o acoso no es algo nuevo (Tarapdar y Kellett, 2012); en cambio, el ciberacoso ha crecido y sus consecuencias basadas en la agresividad con las que las diferentes maneras de hacer ciberacoso (Sameer y Patchin, 2008), son emocionalmente devastantes e inclusive puede llegar a problemas mortales (Navarro *et al.*, 2012; Campbell, 2005) y más para los jóvenes que son los que utilizan de manera frecuente el internet y las redes sociales (Casas *et al.*, 2013).

Los trabajos de investigación que se han realizado se enfocan en detectar el ciberacoso al igual como detectan el acoso basándose en reportes de acusaciones de las víctimas hacia los agresores. Pero esto se da cuando ya se tiene el caso de ciberacoso, cuando la víctima ya habla y menciona su caso y con ello alertar a los profesores, padres de familia y/o adultos responsables de la víctima en sí.

Es por esto que se busca encontrar una manera de poder detectar casos de ciberacoso de manera automática en redes sociales, con herramientas dentro del área de las tecnologías de información que por cuestiones de datos son más adecuados de encontrar este tipo de acoso en estas plataformas sociales. Apoyar a que se detecte de una manera más ágil y dinámica y con ello ayudar al estudio del ciberacoso para lograr prevenir estos casos a tiempo y no cuando sea demasiado tarde. Por ejemplo, que en la red social de la víctima, automáticamente se coloque una marca de color rojo cuando el mensaje que está recibiendo es agresivo, solamente para que esto suceda es porque ya existe una frecuencia de mensajes agresivos de parte del agresor. Así el receptor toma la decisión de leerlo o no, inclusive los padres de familia tienen un apoyo visual del tipo de mensajes que está recibiendo su hijo o hija según sea el caso.

El análisis de sentimiento es una técnica que se ha convertido en un área de investigación cada vez más relevante debido al peso o influencia de los comentarios que expresan los usuarios por algún tipo de tema, producto o tópico en específico y estas expresiones se plasman de manera convincente en las redes sociales.

1.3 Descripción del problema

En esta investigación el objeto de estudio son las redes sociales a través de los mensajes que se publican en este medio. De estos mensajes se crea un repositorio en donde, además de conocer el contenido de los mensajes, se conoce a los emisores y receptores de éstos en una secuencia que permite realizar un seguimiento.

Las variables independientes son los textos de los mensajes y el conjunto emisor-receptor comprometidos en el envío de mensajes. La variable dependiente es la identificación de casos de ciberacoso.

El problema se puede dividir en tres subproblemas relacionados: Identificación de texto agresivo, identificación de acosadores y víctimas potenciales e identificación final de acosadores y víctimas y detectar los posibles casos de ciberacoso. Para la primera parte del problema, la variable de entrada es el texto de los mensajes y la variable de salida es el nivel de agresividad de cada mensaje. Para la segunda parte, la variable de entrada es el conjunto de emisor-receptor y la salida es la identificación de posibles agresores y víctimas. Para la tercera parte del problema, la variable de entrada son los posibles agresores y víctimas y los niveles de agresividad de los mensajes en una línea de tiempo y la salida es la detección de los casos de ciberacoso.

1.4 Preguntas de investigación

La falta de supervisión en el internet, específicamente en las redes sociales, así como las víctimas de ciberacoso que han terminado con su vida, han generado cuestionamientos importantes que motivan el desarrollo del presente proyecto de investigación, por lo cual se generan las siguientes preguntas:

- ¿Es posible detectar el ciberacoso en las redes sociales a través del análisis de sentimiento?
- ¿ Es posible generar un valor, un apoyo numérico y que con esto sea sencillo tomar

una decisión que ayude a detectar agresividad en textos de una manera autómatica?

- ¿Se puede detectar de manera automática el ciberacoso a través de los textos?

1.5 Hipótesis

Un ataque de ciberacoso pueda ser detectado de manera automática a través de técnicas y/o herramientas dentro de las tecnologías de información.

1.5.1 Hipótesis particulares

- Se puede generar un nivel de agresividad el cual aportará a la detección de un ataque de ciberacoso de manera automática.
- La detección de ciberacoso puede tratarse como una subtarea del análisis de sentimiento.
- A través de los tiempos en que se reciben o envían mensajes se pueda detectar de manera automática el ciberacoso.

1.6 Objetivos

Contribuir al desarrollo de un enfoque que permita avanzar en la detección del ciberacoso de manera automática utilizando técnicas y/o herramientas dentro del área de las tecnologías de información como el análisis de sentimiento mediante la clasificación de texto.

1.6.1 Objetivos Particulares

- Identificar palabras agresivas las cuales apoyarán a la identificación de un nivel de agresividad.

- Adjudicar un nivel de agresividad en los mensajes, para así percibir la agresividad que se encuentra en el texto y poder identificar de una manera más oportuna un ciberacoso.
- Determinar a través de la frecuencia de envío de mensajes, si estos forman parte de un ciberacoso.

1.6.2 Contribución

La contribución de esta tesis se basa en:

- Definir un nivel de agresividad el cual identifica de una manera adecuada la agresividad de un mensaje obtenido de una red social y con ello apoya en la toma de decisión si el mensaje es agresivo o no, ya que permite decir qué tanto es lo agresivo del mensaje.
- Identificar a los participantes de las conversaciones que contengan los mensajes agresivos obtenidos de manera dinámica con el nivel de agresividad generado.
- Determinar la frecuencia con la que se envían los mensajes agresivos de los posibles agresores hacia las posibles víctimas, considerando que ya se cuenta con esta información.

Por lo tanto teniendo estos factores: la detección de mensajes agresivos, la detección de los participantes de las conversaciones (el agresor y receptor) y la frecuencia de este envío de mensajes por parte de los participantes, se obtiene la detección del ciberacoso; como se ha mencionado anteriormente, la definición del ciberacoso es un acto agresivo (envío de mensajes agresivos) hacia una víctima (participantes) utilizando vías electrónicas y medios de comunicación electrónica (redes sociales) de manera consecutiva (frecuencia).

1.7 Estructura de la tesis

El presente trabajo de investigación se encuentra estructurado de la siguiente manera: en el Capítulo 2 se describe el marco teórico el cual, contiene los conceptos necesarios de las diferentes áreas que contribuyen al desarrollo de este trabajo como lo es el ciberacoso, lógica difusa, análisis de sentimiento, minería de datos, minería de texto y recuperación de información. En el Capítulo 3, se detallan las investigaciones que se encuentran en el estado del arte.

En el Capítulo 4, se detalla la metodología estratégica para la detección del ciberacoso. En el Capítulo 5, se describen los experimentos que se llevaron a cabo y los resultados obtenidos. En el Capítulo 6, se describen las conclusiones de la investigación y se comenta de una manera breve el trabajo futuro que queda por realizar.

Capítulo 2

Marco Teórico

Sólo porque algo no haga lo que era previsto
no quiere decir que sea inútil el esfuerzo.

Thomas Alva Edison

Este capítulo presenta las áreas de conocimiento necesarias para comprender la investigación realizada. En esta sección se presenta la definición de ciberacoso el cual apoya a explicar las características con las que se formuló la manera de poder resolver el problema de investigación y poder entablar el estudio con las redes sociales.

De las redes sociales es donde se obtienen los mensajes que se necesitan para el estudio, por ello se requiere conocer cuales son los mensajes que nos interesan para la investigación.

Existen varias herramientas que se utilizan para obtener la detección de mensajes agresivos en las redes sociales, por ello, se detallan las herramientas para obtener la información requerida como por ejemplo las áreas en las que nos apoyaron para realizar este trabajo como el análisis de sentimiento y minería de datos.

2.1 Ciberacoso

El problema del acoso tradicional no es considerado como un problema social nuevo, se ha estudiado su comportamiento y sus definiciones. Se considera el acoso como un abuso en el trato hacia una o varias personas utilizando violencia de diferente índole, los cuales, pueden llegar a ser hasta golpes; solo que no cualquier confrontación puede ser considerada acoso; para que este sea acoso se requiere de al menos de dos personas, el agresor y el agredido o víctima, aunque claro se ha estudiado que en un caso de acoso se pueden llegar a involucrar de una manera indirecta más de dos individuos (Tarapdar y Kellett, 2012; Campbell, 2005; Harris y Petrie, 2002).

Las agresiones indirectas, a través de una tercera persona incluyendo un daño como agresión relacional, el cual se basa en afectar las relaciones de un individuo con sus contactos, es también una manera de realizar acoso que ha ido en ascenso. Este tipo de acoso ha abierto una variante para realizar daño intencional, con herramientas tecnológicas apropiadas y disponibles y con ello originando el ciberacoso (Smith y Collage, 2006).

2.1.1 Definición de ciberacoso

La definición de ciberacoso considerada para nuestra investigación es la siguiente: El ciberacoso es un acto premeditado, agresivo, que se realiza a través de dispositivos tecnológicos, tales como: teléfonos inteligentes, tabletas, computadoras portátiles, de escritorio; dirigido hacia un grupo de personas o a una persona considerada víctima o víctimas, de manera recurrente, originado por un agresor (Smith, 1999; Espelage y Swearer, 2003).

2.1.1.1 Características del ciberacoso

El ciberacoso cuenta con ciertas características peculiares como lo son: El desconocer la identidad del agresor, el nivel de audiencia de hasta donde se puede llegar el abuso ocasionado por el agresor hacia la víctima, el poder que posee el agresor al contar con cierta estimulación para generar un ataque, ya sea para ridiculizar o herir al agredido en

toda una red social (Juvonen y Gross, 2008; Suler, 2004; Dinakar *et al.*, 2012)

El acoso puede ser generado gracias a numerosas tecnologías que se han desarrollado (Mendoza, 2012; Slonje *et al.*, 2012; K Jowalski R, 2010) tales como:

- Mensajes Instantáneos: Los mensajes instantáneos en tiempo real de una persona a otra u a otras, como en un grupo de contactos, ejemplos de estas herramientas son: Messenger, Whatsapp, Skype, etc.
- Mensajes de texto *(SMS o MMS Technology [short message service o multimedia messaging service])*: A diferencia de los mensajes Instantáneos estos no son en tiempo real, se realizán por medio de un servicio telefónico móvil.
- Correo electrónico: es un servicio de mensajería que utiliza como metáfora el correo convencional. Se encuentra conformado por un emisor, receptor y su mensaje que conforma el cuerpo del correo.
- Chats: son espacios en línea en donde se permiten a grupos de personas tener acceso y conocerse gracias al intercambio de información que comparten entre sí: ejemplo: www.neltingo.com; en este sitio puedes usar tu propia identidad o cambiarla al gusto.
- Blogs: son páginas web que sirven como bitácoras o diarios personales en donde usuarios que pueden leer estos mensajes pueden emitir su comentario o compartir experiencias al respecto dejando un mensaje a quien origina el blog.
- Sexting: es una palabra compuesta de sexo y texto, este es un mensaje con contenido sexual o una fotografía sexualmente explícita enviada por mensaje de texto, usando SMS *(short message service)* y/o tecnología MMS *(multimedia messaging service)*.
- Juegos de internet: Son juegos de consola como los son el X-box®, apps de android ®o IOS ®y en general juegos de PC, los juegos de internet que se encuentran considerados como *live action* son los que se puede interactuar entre los jugadores y agredirse entre ellos mismos durante el juego.
- Redes sociales como lo son: *Path, Oink, Bebo, Stamped, Linkedin, Twitter, Facebook, MySpace, Livejournal, Friendter, Nexopia, Xuga, Xanga, Impee.* Son una ventana al mundo juvenil, ya que permiten identificar las actividades de los usuarios en tiempo

real. Para estar incluido en las redes sociales se requiere cierta edad, usualmente son gratuitas e invitan a los usuarios a colocar información personal como perfiles y fotografías, lo que a su vez generan vulnerabilidad al ser aprovechadas esta información para generar y enviar mensajes agresivos a los que conforman estas redes.

Debido a la definición de ciberacoso que se mencionó anteriormente y a los problemas que estos conlleva se considera trabajar para esta investigación la búsqueda de mensajes agresivos, dirigidos hacia un usuario, de manera frecuente por medio de una red social.

2.2 Procesamiento de información para la búsqueda de mensajes y usuarios agresivos

Las redes sociales cuentan con una extensa cantidad de mensajes publicados, es por ello que para encontrar la "aguja en el pajar" refiriéndose a los mensajes agresivos que se necesitan para la investigación se requieren de varios procesos.

Entre estos procesos se encuentra el filtrado de los mensajes para seleccionar los mensajes que son considerados agresivos y los que no son agresivos, además de eliminar faltas de ortografía y palabras escritas de manera incorrecta.

2.2.1 Recuperación de información

La recuperación de información o *Information retrieval* es un proceso para obtener información destacada, dentro de una coleción de documentos que hablan sobre el tema en cuestión. Esta información obtenida se basa en los requerimientos solicitados por un usuario (Baeza-Yates *et al.*, 1999; Manning *et al.*, 2008).

Para nuestra investigación, los documentos se encuentran conformados por los mensajes de texto que se colocan en la red social, la colección de documentos se refiere a la base de datos con los mensajes obtenidos de la red y teniendo en cuenta esta información, se obtienen los mensajes relevantes, utilizando los criterios fundamentales que nos

proporciona la recuperación de información.

Debido a estas definiciones y para poder obtener la relevancia de los documentos, se considera para la metodología de nuestra investigación los documentos que vienen siendo los mensajes, por lo que se requiere de la recuperación de la información con sus criterios fundamentales.

Los criterios fundamentales que se encuentran dentro de la recuperación de información son la tarea de usuario y la vista lógica de los documentos.

2.2.1.1 Tarea del Usuario

La tarea del usuario consta de lo que el usuario tiene que realizar para poder obtener la información que requiere. Entre lo que solicita el usuario y la tarea del mismo se unen por medio de un lenguaje que es considerado implícito para el sistema. Por ende el sistema permite la recomendación o la recuperación de documentos que contienen la información solicitada (M.F, 2002; Baeza-Yates *et al.*, 1999). Para nuestro trabajo de investigación esta tarea es la de identificar la agresividad por medio de las palabras agresivas de nuestro repositorio. El cual se complementa con la vista lógica de documentos.

2.2.1.2 Vista lógica de los documentos

La navegación es un camino metodológico por la colección de documentos en lo que se define lo que se requiere de información (Baeza-Yates *et al.*, 1999).

Cuando el documento es considerado grande (millones de documentos) es adecuado resumirlo a una lista de palabras clave del mismo texto. Con ello se obtiene la primera representación denominada *texto completo* (Tolosa y Bordignon, 2008; M.F, 2002). El repositorio de nuestra investigación es considerado grande, debido a los millones de mensajes que se tienen que revisar para detectar los mensajes que se requieren y por lo que se resume a una lista de palabras que nos apoyan a encontrar la solicitud del usuario que es encontrar agresividad en los mensajes de texto, de un punto de vista lógico para el sistema.

Un documento representado dentro de sus índices es considerado vista lógica (M.F, 2002).

La recuperación de información se encuentra relacionada con la metodología de esta investigacón de tesis gracias a los modelos adecuados que maneja para obtener la relevancia de los documentos y con ello poder lograr obtener los mensajes que se requieren de los textos relevantes, con las palabras que se solicitan, siendo estos los mensajes que contienen palabras agresivas.

2.2.2 Análisis de sentimiento

Antes de que existiera el internet, la manera en la que se buscaban recomendaciones para poder tomar una decision era con base en las opiniones de amigos, conocidos y familiares.

Hoy en día, el internet ha apoyado en el proceso de la toma de decisiones, por ejemplo, para escoger cual auto o cual casa es más adecuado, se cuentan con sitios web que nos apoyan con la posibilidad de facilitarnos el conocer las recomendaciones de otras personas basadas en sus opiniones y experiencias, estas personas no tienen que ser conocidos de los usuarios(Pang y Lee, 2008).

El procedimiento para la toma de decisiones es considerado una información clave que se encuentra dentro del área del análisis de sentimiento; este procedimiento consiste en comparar las opiniones de varias personas consideradas expertas en su área junto con personas que no lo son (Pang y Lee, 2008; Feldman, 2013).

El análisis de sentimiento o minería de opinión apoya a revelar opiniones relacionadas sobre un tema en específico. Este tema de investigación se ha vuelto muy frecuente en el área de las tecnologías de información; una fuente disponible de estas de opiniones basadas en sentimientos se encuentran en redes sociales como *Facebook, Twitter, blogs* y foros de usuarios (Feldman, 2013).

El área del análisis de sentimiento se encuentran en campos determinados, estos campos se centran en ciertos problemas las cuales son (Liu, 2010, 2012; Pang y Lee, 2008):

- Análisis de sentimiento a nivel de documentos;

- Análisis de sentimiento a nivel de enunciados;
- Aspectos basados en el análisis de sentimiento;
- Comparación del análisis de sentimiento; y,
- Adquisición del sentimiento a través del léxico.

Dentro de esta investigación se trabajará con dos áreas el análisis de sentimiento a nivel de documentos y el análisis de sentimiento a nivel de enunciados, las cuales se definirán a continuación:

2.2.2.1 Análisis de sentimiento a nivel de documentos

El análisis de sentimiento a nivel de documentos es el más sencillo de los niveles que forman parte del estudio del análisis de sentimiento, este a su vez considera que la opinión del autor del documento sobre un tema importante se encuentra manifestada en el documento. El análisis de sentimiento a nivel de documentos se encuentra conformado por el aprendizaje supervisado y no supervisado. El aprendizaje supervisado presenta un conjunto finito de clases, en donde puede ser clasificado el documento. Cada una de las clases que presenta el conjunto se analizan con los datos entrenados que están disponibles para las clases en cuestión. Las clases que son para cuando un caso es sencillo o simple, son positivo y negativo. Una clase neutral o la consideración del manejo de una escala en donde se le deberá de colocar a cada documento (como el sistema de cinco estrellas que utiliza *Netflix(www.netflix.com)*) son consideradas extensiones simples en el análisis de documentos. El sistema de este análisis consiste en aprender de un modelo de organización a través de utilizar un algoritmo común de clasificación como pueden ser máquinas de soporte vectorial, teorema de Bayes o regresión lógica cuando se le aporta a un sistema datos entrenados. En las diversas clases que especificarán el sentimiento se utilizará este tipo de clasificaciones para agregar una etiqueta a la nueva documentación. Cuando es un valor numérico (en algunos un rango finito), la etiqueta que se le asigna al documento, entonces, la regresión puede ser utilizada para predecir el valor que puede ser asignado al documento (Feldman, 2013).

Investigaciones han demostrado que una buena proximidad de pertenencia asignada al documento es alcanzada cuando cada documento está representado por una simple bolsa de palabras *(bag of words)*. El llamado <<saco de palabras>> *(bag-of-words)*, es el modelo más común, el cual simplemente toma las palabras diferentes en el documento (Hotho *et al.*, 2005). A cada palabra, posteriormente se le puede asignar un peso; lo cual se puede llevar a cabo de diferentes formas.

Estas formas más avanzadas que se utilizan dentro del área del análisis de sentimiento son: la frecuencia de ocurrencia del término en la colección de documentos *(TFIDF, term frequency Inverse-document frequency*, comúnmente utilizado en el campo de la minería de datos (Wong y Yao, 1992)), información que se toma como parte del texto *(POS,Part of Speech information)*, léxico de sentimiento y estructuras de análisis sintáctico (Pang *et al.*, 2002).

El aprendizaje no supervisado se encuentra basado en la determinación de la orientación semántica *(SO, semantic orientation)* de frases específicas que se encuentran en el documento. Si el promedio de la orientación semántica de estas frases está por encima de un umbral predefinido el documento será clasificado como positivo, de lo contrario, se considerará negativo.

Hay dos enfoques principales para la selección de las frases: un conjunto de información de parte del texto *(POS)* predefinido puede ser utilizado para seleccionar estas frases(Turney, 2002) o un léxico de palabras de sentimiento (Taboada *et al.*, 2011). El enfoque basado en utilizar lexicones es con el que se va a trabajar en esta tesis.

El uso de éstos es común en el análisis de sentimiento para el idioma inglés.

Los lexicones que se utilizan para esta investigación son:

`noswearing.com` es un sitio que almacena contribuciones por parte de la comunidad anglosajona de palabras ofensivas así como su significado. Es la misma comunidad quienes aportan las palabras a este sitio, la lista tiene como beneficio que se encuentra compuesta por el *slang* y por palabras que van surgiendo con el tiempo, esto quiere decir que se encuentra actualizado (Sood *et al.*, 2012).

ANEW , «*Affective Norms for English Words*», se formó por los participantes que evaluaron su reacción a un conjunto de 1034 palabras con respecto a tres estándares semánticos diferenciales de bueno-malo (valencia psicológica), activo-pasivo (motivación) y fuerte-débil (dominio). Se utiliza una escala del uno al nueve, donde uno es lo menor en la escala de cada uno de los estándares y nueve el mayor en cada uno de los estándares (Dodds y Danforth, 2010).

SentiWordNet es una herramienta léxica para la minería de opinión. Esta herramienta utiliza la base de datos de WordNet, una base de datos que contiene las palabras en inglés. WordNet contiene nombres, verbos, adjetivos y adverbios agrupados en conjuntos de sinónimos cognitivos, llamados <<*synsets*>>. Cada *synset* expresa un concepto distinto. Los *synsets* están vinculados entre sí por medio de las relaciones conceptuales semántico y léxico. SentiWordNet asigna a cada *synset* de WordNet tres clasificaciones con respecto a la confianza, la negatividad, la positividad y la objetividad.

Cada *synset* se asocia a tres valores numéricos Pos(s), Neg(s) y Obj(s) que indican los términos positivos, negativos u objetivas (o neutros) están contenidas en cada *synset*; cada valor está dentro del intervalo [0.0, y 1.0] y la suma de los tres valores asociados es necesariamente 1.0. Esto significa que cada *synset* tiene un valor distinto de cero en al menos una de las categorías (Ventura de Souza, 2011).

2.2.2.2 Análisis de sentimiento a nivel de enunciados

Un documento por más simple que parezca puede contener diversos puntos de vistas sobre los mismos temas. En el momento que se está buscando un punto de vista más especifico de las expresadas en el documento y que tratan sobre el tema es adecuado utilizar un nivel basado en enunciados (Feldman, 2013).

El nivel basado en enunciados comienza por considerar que se cuenta con la esencia del enunciado a analizar. Se continua considerando que se cuenta con un punto de vista u opinión sencilla de cada enunciado. Esta consideración se puede partir en dos frases en la cual cada frase contiene únicamente una opinión.

El análisis de la polaridad del enunciado se realiza después de conocer si los enunciados son objetivos o subjetivos. Debido a que los enunciados que son considerados solamente para el análisis son los enunciados subjetivos (Feldman, 2013)..

Otro tipo de orientaciones estudian los enunciados subjetivos, debido a su dificultad. Para lograr la clasificación de los enunciados en clases, una gran parte de estas orientaciones, emplean enfoques supervisados (Yu y Hatzivassiloglou, 2003).

Trabajar con diferentes tipos de enunciados por métodos distintos es lo que consideran en las investigaciones actuales, dentro de estas investigaciones se consideran métodos para enunciados que contienen condicionantes, preguntas y sarcasmo (Narayanan *et al.*, 2009). El sarcasmo es difícil de encontrar y existe principalmente en varios contextos como el político (Tsur *et al.*, 2010).

El enfoque basado en utilizar lexicones como *ANEW,SentiWordNet, NoSwearing* y junto con el contador de frecuencias de las palabras agresivas encontradas en los mensajes obtenidos de la red social que cuentan con la solicitud de información por parte del usuario, son las herramientas que se utilizan en esta área del análisis de sentimiento y es lo que se utilizan en la metodología con la que se va a trabajar en esta tesis.

2.2.3 Lógica difusa

La lógica difusa o *fuzzy logic* fue propuesto por Lofti A. Zadeh, en donde realizó una adecuación a la teoría de conjuntos (Cañellas y Brage, 2006; DNegri y De Vito, 2006).

En la teoría de conjuntos *(crisp sets)* se consideran dos valores,si un elemento forma parte de un conjunto o no forma parte del conjunto. En cambio en la lógica difusa se cuentan con conjuntos difusos los cuales estos representan información más amplia, que acepta la pertenencia de elementos en un determinado conjunto que asemeja difícil de determinar (DNegri y De Vito, 2006).

La altura de un grupo de personas es un ejemplo frecuente para explicar la lógica difusa; en donde, se define un universo de los posibles alturas de los individuos. En la figura 2.1 se representa de manera gráfica el universo de las alturas de un conjunto de personas,

en donde puede lograr definir los conjuntos de quienes son personas altas y quienes no. Además que se perciben los conjuntos difusos y sus funciones de pertenencia.

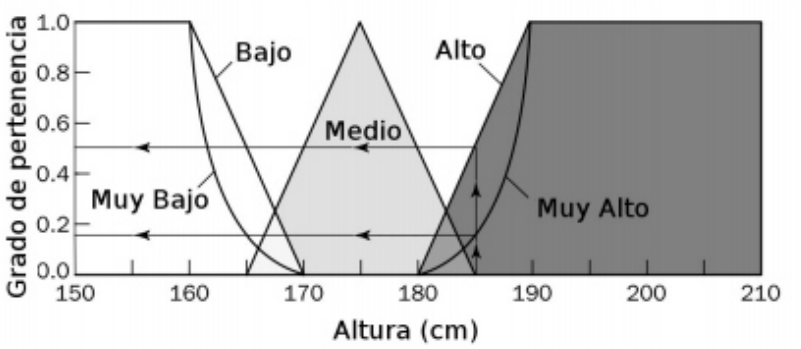

Figura 2.1: Figura que ejemplifica el universo de alturas en una población de individuos (González Morcillo, 2012)

Los conjuntos difusos no cuentan con una definición precisa de pertenencia, ya que el grado o función de pertenencia es el que demuestra el valor de que tanto pertenece el elemento al conjunto difuso. En la imagen 2.2 se representa el rango de valores que este grado de pertenencia puede obtener, como se observa es entre los valores desde cero hasta uno, en donde el cero significa que no pertenece al conjunto y entre que el valor se acerque más al uno esto quiere decir que tiene más pertenencia.

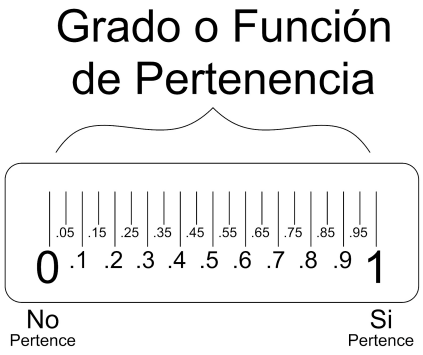

Figura 2.2: Figura que ejemplifica el rango de valores que puede tomar el grado o función de pertenencia

Siendo este el concepto que remplaza la polaridad de si/no, blanco/negro, considerando estos últimos como conceptos intrínsecos y dependientes del dominio del conjunto. Por medio de la teoría de los conjuntos difusos que forman parte de la lógica difusa se puede imitar la forma en que los humanos toman decisiones, como por ejemplo: si mencionamos "Petra terminará la tesis en unos pocos meses", esto se considera difuso (ambiguo, pero proporciona información), sin embargo al mencionar: "Petra terminará la tesis alguna vez", esto sería considerado vago (inexacto y sin proporcionar información). En comparación a la lógica de Boole (1854),en la lógica difusa se consideran las opciones de verdad y falso, además de sus valores intermedios (DNegri y De Vito, 2006).

Por lo tanto se puede considerar que la lógica difusa o borrosa es una opción a la lógica fundamentada en conjuntos discretos en donde su finalidad es conocer si alguien o algo se encuentra en un conjunto o no determinado según obedezca a condiciones establecidas (en nuestro trabajo de tesis a que si un comentario es agresivo o no).

Por tanto la lógica difusa es un método de razonamiento que se considera en diferentes grados de valores de veracidad o que al igual para tomar decisiones a las categorías limitadas durante la resolución de problemas. (Klir *et al.*, 1997; Mendel, 2001; Cañellas y Brage, 2006).

Como ejemplos en el area de lógica difusa, el estudio de procesamientos complejos, confusos o desordenados la lógica difusa a realizado cambios además en otras áreas. La lógica difusa, se considera pertinente para considerarse en el manejo de situaciones indeterminadas y complejas, en cambio los algoritmos son considerados para dar razón de procesos determinados, por lo que, se encuentran aislados de contextos complicados y confusos. En el área de las programación de los sistemas expertos, es tomada en cuenta como una herramienta esencial, con beneficio en la rama de la inteligencia artificial. Las aplicaciones de la lógica difusa se enfocan en las áreas que solicitan elementalmente de control, revisión de toma de decisiones o reconocimiento de patrones (Cañellas y Brage, 2006).

En las áreas de economía y de finanzas, la lógica difusa ha tomado fondo, así como en medicina, por medio de sistemas expertos que colaboran a la toma de decisiones en diagnósticos médicos (McNeill y Freiberger, 1994; Brassler y Homburg, 1996).

Cabe decir que la mayoría de los trabajos que aúnan educación y lógica difusa corresponde a estudios sobre la propia enseñanza de la teoría de los conjuntos difusos *(fuzzy sets)* en las escuelas de ingeniería, de robótica y de tecnología. La lógica difusa apoya a estudios que avalan la enseñanza de la teoría de los conjuntos difusos, correspondiendo a la rama de la educación en las escuelas de ingeniería, de robótica y de tecnología (Cañellas y Brage, 2006).

2.2.4 Sistemas difusos

Un sistema difuso funciona por medio de un conjunto de entradas y proporciona un resultado generado por medio del motor de inferencia que se utiliza en base a reglas difusas. La esencia de un sistema difuso se encuentra formado por los conjuntos difusos, las variables lingüisticas y las funciones de pertenencia. Un conjunto difuso genera un valor de pertenencia para cada elemento del sistema; por ejemplo en el conjunto Edad = $\{(32, 0.1), (55, 0.5), \ldots (97, 0.9)\}$, el elemento 97 años de edad tiene un valor de pertenencia de 0.9. Existen varias funciones de pertenencia, las triangulares y trapezoidales son las que se frecuentan.

$$triangular(z, a, b, c) = max\left(min\left(\frac{(z-a)}{(b-a+\epsilon)}, \frac{(c-z)}{(c-b+\epsilon)}\right), 0\right), \tag{2.1}$$

$$trapezoidal(z, a, b, c, d) = max\left(min\left(min\left(\frac{(z-a)}{(b-a+\epsilon)}, 1\right), \frac{(d-z)}{(d-c+\epsilon)}\right), 0\right). \tag{2.2}$$

En las ecuaciones, 2.1 and 2.2, z es un elemento dado del conjunto difuso y $\epsilon = 0.000001$ se utiliza para proteger la división. Los parámetros $a, b, c,$ y d se ilustran en las figuras 2.3 y 2.4.

Una *variable lingüistica* es una variable asociada con otra variable x y no considera valores numéricos sino valores lingüisticos. Por ejemplo:

$$\text{Comentario} = \{\text{“Muy Corto”}, \text{“Corto”}, \text{“Moderado”}, \text{“Largo”}, \text{“Muy Largo”}\}.$$

Cada valor lingüistico esta relacionado con un conjunto difuso o una función de

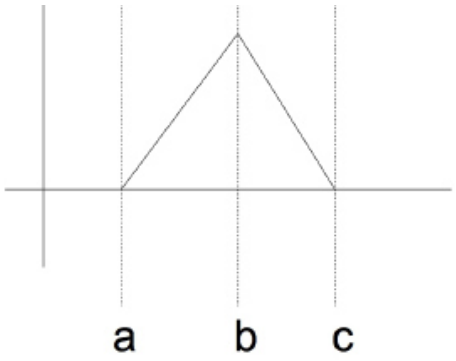

Figura 2.3: Parámetros y función de pertenencia triangular

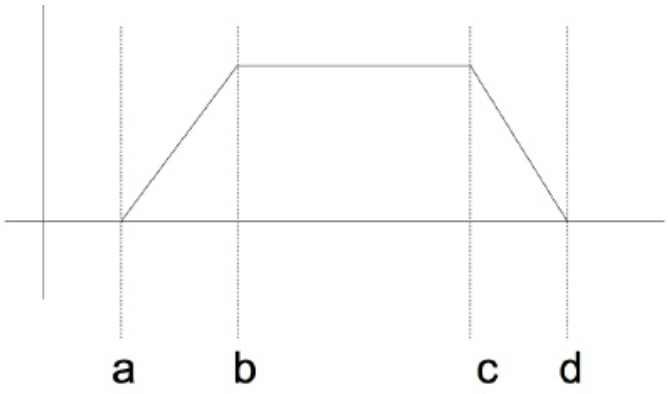

Figura 2.4: Parámetros de función de pertenencia trapezoidal

pertenencia. El rango de operación esta definido por cada variable lingüistica y se encuentra particionado según los valores lingüisticos que se utilizan. Normalmente se utiliza un numero non para las particiones, siendo 3 o 5 lo más común en nuestro caso fueron 5, Figura 2.5.

Respecto al sistema de inferencia y en base a las reglas de los conjuntos difusos, en primer lugar, la base de reglas difusas es una matriz *(Mt)* donde cada fila es una regla y cada columna es un entero que representa el valor lingüistico asociado con la regla de antecedentes y consecuentes. Después considerando un intervalo para cada variable lingüistica, una función $tipo1(x, n)$ puede definirse de la siguiente manera:

Figura 2.5: a) Cinco conjuntos difusos: 1- Muy Corto, 2- Corto, 3 - Moderado, 4 -Largo, 5 -Muy Largo.

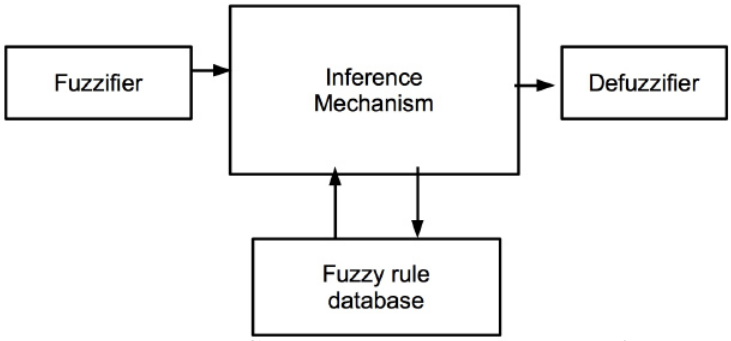

Figura 2.6: Componentes de un sistema difuso

$$tipo1(x,n) = \begin{cases} trapezoidal(x,0,0,0.1666,0.3333) & \text{if } n=1 \\ triangular(x,0.1666,0.3333,0.5) & \text{if } n=2 \\ triangular(x,0.3333,0.5,0.6666) & \text{if } n=3 \\ triangular(x,0.5,0.6666,0.8333) & \text{if } n=4 \\ trapezoidal(x,0.6666,0.8333,1,1) & \text{if } n=5, \end{cases} \tag{2.3}$$

donde x es un valor de entrada y n corresponde especificamente a la función de pertenencia de la variable lingüistica. Considerando un sistema difuso de tres entradas un mecanismo de inferencia *max-min* se define como :

$$I(p) = \min(tipo1(x_1, Mt(p)), tipo1(x_2, Mt(p)), tipo1(x_3, Mt(p))) \tag{2.4}$$

donde p es el numero de la regla difusa,Mt es el sistema disfuso en el que se basa I es la inferencia calculada.

Para finalizar, la salida puede ser determinada por el desfuzzificador centralizado y alto, como se muestra a continuación:

$$yh = \sum_{p=1}^{R}(I(p) * c_g(Mt(p)))/\sum_{p=1}^{R} I(p) \tag{2.5}$$

```
{Version centralizada}
{desfuzificador centralizado}
sum1=0;sum2=0.00000001;
for dy=0,0.01,1{from 0 to 1, steps from 0.1}
    ss=0;
    for r=1,NTRules
        n = Mt(R,I+0);
        if n>0
            V = PARAM(Mt(R,I+0),:); { Extracto de parametro del conjunto
             difuso involucrado en la regla R}
        fin de la condici\'on
        mf = Type1FS(dy,n,V);
        ss = max(ss,min(mf,I(R)));
    fin del ciclo
    sum1 = sum1+ss*dy;
    sum2 = sum2+ss;

fin del ciclo

y=sum1/sum2;
```

Figura 2.7: Versión centralizada del defuzificador centralizado

donde c_g es un centro de masa precalculado de cada función de pertenencia de la salida y r es el numero de reglas. Mt establece el valor correcto dependiendo de

las reglas difusas.

La lógica difusa en nuestra investigación fue parte de la experimentación para la comparación de métodos que se consideran fáctibles para la detección de agresividad en los mensajes de texto.

2.2.5 Minería de datos

Se consideran como el florecimiento de tecnologías nacientes al *(Data-mining)* o la minería de datos, la minería de texto o *(textmining)* minería textual las cuales apoyan al estudio del conocimiento que se incluya en los datos almacenados (Liu, 2007).

La minería de datos se encuentra establecida como el conocimiento que se encuentra en base a patrones que pueden ser detectados en datos estructurados, como por ejemplo en la base de datos relacionales. Es por lo anterior descrito que se le considera usualmente a la minería de datos como *Knowledge-Discovery in Databases (KDD)* , considerando este como el proceso que consta del reconocimiento de información útil, inclusive utilizando patrones adecuados, en bases de datos (Fayyad *et al.*, 1996). Siendo este el proceso que utilizamos en nuestra metodología para esta investigación.

Los datos a los que hace mención la definición son sucesos que han sido registrados en una base de datos, por tanto los patrones son expresiones de un lenguaje determinado, utilizados para asignar subconjuntos de los hechos registrados en la base de datos. Esto significa que estas expresiones requieren de la técnica o método utilizado para el análisis de los datos. Por ejemplo, si la técnica es la clasificación aglomerativa, los patrones suelen ser grupos *(clústers)* o particiones (Vairinhos, 2003).

Los patrones deben ser comprendidos por los seres humanos. Tienen que ser sencillos y entendibles, que sean de una fácil comprensión en el idioma coloquial de

los seres humanos.

Considerando el uso de información anterior y la que se requiere utilizar dentro de la toma de decisiones es adecuado considerar que los patrones deben ser válidos, útiles, novedosos, repentinos y distintos de los valores esperados esto debido a su trascendencia en el uso de los datos al momento de su descubrimiento ya que con ellos se tiene significado para otros datos o son considerados para procesos de tomas de decisión.

Además definir patrones en la información, la minería de datos puede ser utilizada para aumentar recursos, disminuir costos o hacer las dos funciones y es un proceso el cual consiste de varias fases o pasos en el análisis de información el cual parte desde diferentes puntos de vista.(Fayyad *et al.*, 1996). El proceso de minería de datos, se encuentra en la imagen 2.8

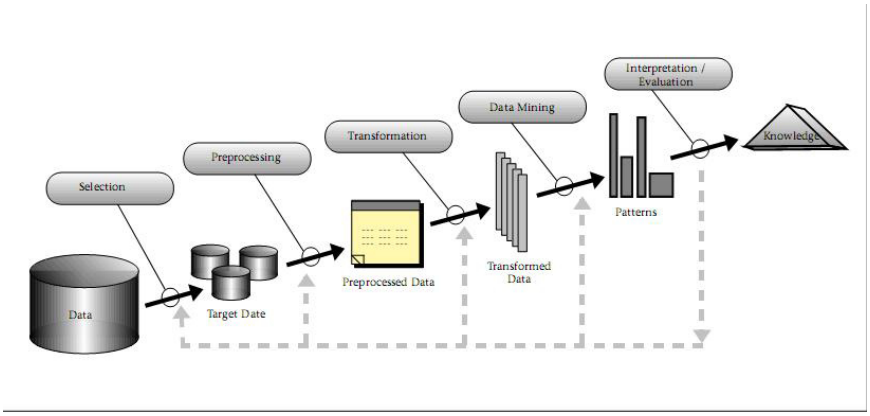

Figura 2.8: Proceso de la minería de datos (Fayyad *et al.*, 1996)

2.2.6 Minería de texto

La minería de texto está dirigida hacia el origen de conocimiento a partir de información no-estructurada en un lenguaje natural. Esta información se encuentra almacenada en bases de datos textuales las cuales reconocen patrones de manera automática que son importantes y no banales de conocimiento en los documen-

tos (Feldman y Dagan, 1995; Feldman y Sanger, 2007). A esto se le define como Knowledge-Discovery in Text (KDT) (Fayyad *et al.*, 1996; Galvez, 2012).

Además es considerado como un conflicto en la clasificación de texto o de agrupamiento dependiendo de algunas semejanzas con la información para analizar, de misma manera el área de la minería de texto se encuentra relacionada con otras como la recuperación de información, análisis de texto, procesamiento de lenguaje natural, clustering, aprendizaje máquina y minería de datos (Berry y Castellanos, 2004; Tan *et al.*, 1999).

2.3 Resumen

Los temas que se mostraron en este cápitulo, es sobre la teoría que sustenta la investigación llevada a cabo en esta tesis. Desde lo que es ciberacoso, redes sociales y las áreas que apoyan a encontrar agresividad en los mensajes, como análisis de sentimiento, minería de datos y minería de texto. Además como la recuperación de información, lógica difusa y lexicones que ayudan para llegar a detectar el ciberacoso.

Capítulo 3

Estado del arte

Si buscas resultados distintos no hagas siempre lo mismo

Albert Einstein

En este capítulo se describen los trabajos de investigación que se encuentran en el estado del arte. Estos trabajos apoyan a fundamentar el camino que ha tomado nuestra investigación.

El capítulo se organiza de la siguiente manera: Primero en la sección 3.1 se presentan las metodolgías para reportar el ciberacoso a través de sistemas sociales y con ello llegar a las agencias de ley. Se continúa con la sección 3.2 con los trabajos de investigación que se encuentran en el estado del arte que hablan sobre detección de ciberacoso en línea. En esta sección se refiere <<en línea>> a que se encuentra en la red, en internet así como en redes sociales, blogs, foros y *chats*. Además en esta sección se encuentra la subsección 3.3 que menciona los trabajos relacionados con la detección de agresores involucrados en el caso del ciberacoso.

Posteriormente se encuentra la subsección 3.4 que habla sobre los trabajos de investigación que desarrollaron aplicaciones para la detección del ciberacoso.

Y por último, se describe las áreas relacionadas con la detección automática

del ciberacoso.

3.1 Metodologías para reportar el ciberacoso a través de sistemas sociales

En el estado del arte, el acoso se ha estudiado en sus diversas características y además desde su punto de vista social (Salmivalli *et al.*, 1996). En el caso del ciberacoso, se han generado estudios en donde se han propuesto diversas metodologías que a través de sistemas sociales ayudan a conocer si existe el ciberacoso como por ejemplo, uno de estos es el SocialFilter (Chen *et al.*, 2012), un sistema de tiempo real que ayuda a padres y educadores a rastrear mensajes de los jóvenes en *Twitter*, especialmente con el fin de detectar si han sido acosados o han intimidado a otros.

El objetivo del sistema de Chen *et al.* (2012), se encuentra en su metodología de las cuatro "I" `( Four Is )` : Identidad del acosador `(Identity of bullies)`, deducir el mensaje agresivo `(Inference the bullying message)`, influenciar en el comportamiento del agresor `(Influence of bully behavior)` e intervenir `(Intervention)`.

Otra metodología con enfoque social es La teoría de la conducta planificada `(the theory of planned behavior (TPB))`, que propone que las actitudes del comportamiento surgen en base a las actitudes que manejan a la persona. Esta investigación se basa en un cuestionario el cual ayuda a predecir la presencia del ciberacoso escolar de los adolescentes obtenida por autoinforme (Walrave y Heirman, 2011).

Una metodología se probó a través de las historias relatadas en un canal social de MTV llamado, *MTV Over the Line?*, donde los usuarios reportan casos potenciales de abusos (Macbeth *et al.*, 2013; Sui, 2015). Esta investigación llevada a cabo por el Instituto Tecnológico de Massachusetts `(MIT)` contempla el uso de scripts para encontrar empatía en las conversaciones para corroborar que su técnica a base de

scripts consta de un patrón que maneja actores, acciones y eventos que describen una historia.

3.2 Detección de ciberacoso en línea

Patchin y Hinduja (2006) definieron el ciberacoso como "un acto malicioso, deliberado y repetitivo causado por medio de un texto electrónico". Desde el punto de vista en la psicología juvenil muestra nueve tipos de ciberacoso que han sido reconocidos por Willard (2007); Patchin y Hinduja (2006); Maher *et al.* (2008).

Estas categorías son:

1. Insulto electrónico *(flooding)*.
2. Suplantación *(masquerade)*.
3. Delatar *(flaming)*.
4. Hostigamiento *(trolling)*.
5. Acoso *(harassment)*.
6. Amenaza cibernética *(cyberstalking)*.
7. Denigrar *(denigration)*.
8. *Outing* Similar a denigrar solo que en este caso el agresor y la víctima tienen una relación en línea o en persona.
9. Excluir *(exclusion)*.

Se han realizado investigaciones basadas en el paradigma de la minería de texto para casos que se han relacionado con la detección de ciberacoso en sus diferentes categorías.

Sin embargo se han realizado pocos estudios que hablen sobre soluciones técnicas, debido a base de datos insuficientes y adecuadas para este tipo de casos; además las cuestiones de privacidad y falta de información pueden ser las razones que describe un ciberacoso (Nadali *et al.*, 2013).

Yin *et al.* (2009) propone un método de aprendizaje supervisado para determinar *posts* acosadores en *chats* y en foros. Para entrenar una máquina de vectores, (SVM)*: Support Vector Machine*, utilizó tres características: contenido específico, sentimiento y contexto, considerándolo como hostigamiento.

Además utilizaron el $n - gramas$, peso $TFIDF$ *(term frequency Inverse-document frequency)* y cuatro palabras frecuentes como punto de referencia (*baseline*). Aunque sus resultados indicaron mejorías sobre el punto de referencia, la información del usuario no la utilizaron del todo. Además ellos utilizaron solamente métodos supervisados. Aún así los métodos no supervisados probaron ser adecuados.

La Investigación de Al-garadi *et al.* (2016) proponen un modelo de detección de ciberacoso el cual también utilizan (SVM). Con ayuda de la red social *Twitter* obtienen los datos utilizando en las siguientes características: red, actividad del usuario en la red, usuario y el contenido del mensaje.

El investigador Chisholm (2006) probó identificar grupos que contienen ciberacoso usando un algoritmo basado en reglas utilizando el mismo conjunto de datos obtenidos en línea que provocaban hostigamiento.

Dinakar *et al.* (2011), describen un método para detectar ciberacoso entre los comentarios que se encuentran en `Youtube.com`, los cuales contenían insultos. En su método utilizan una variedad de clasificadores binarios y multiclases sobre un conjunto de datos etiquetado.

Además aplicaron conocimiento de sentido común para detectar ciberacoso. Utilizando sentido común se ayudó a proveer información sobre las posibles vícitmas de los acosadores, sus emociones y las propiedades y relaciones de las posibles vícit-

mas, lo cual ayudaban a aclarar y contextualizar el lenguaje con el que se hablaban. También utilizaron dos características:

1. Caracteísticas generales que contienen uni-gramas con valor TFIDF [1], el lexicón *Ortony* en su conotación negativa, una lista de palabras profanas y frecuentemente etiquetas de bigramas POS `(Part of Speech information)` [2].

2. Características etiquetadas específicamente.

Su estudio indica que el clasificador binario puede superar el reconocimiento de ciberacoso textual en comparación con los clasificadores multiclase. Sus resultados muestran como el uso de tales características son útiles y pueden conducir a una mejor visualización del problema.

Las limitaciones en su estudio fueron la falta de consideración del diálogo y el grafo de la red social.

Reynolds *et al.* (2011) mejoraron el trabajo realizado por Dinakar *et al.* (2011). Propusieron un método de aprendizaje automático,*(machine learning)*, para detectar ciberacoso de `Formspring.me`, el cual manejaba insultos y hostigamiento.

Aplicaron el numero de palabras consideradas malas *NUM* y la densidad de palabras caracterizadas como malas *NORM*, las cuales fueron consideradas severas si se encontraban en la lista de malas palabras `noswearing.com`. Emplearon su réplica con ejemplos positivos hasta diez veces y la precisión en el rango de los clasificadores fue reportada.

Sus resultados mostraron que el árbol de decisión $C4.5$ y la instancia basada en el aprendizaje pudieron reconocer como verdaderos positivos una precisión de $78.5\,\%$.

Algunos trabajos se han realizado sobre el reconocimiento de usuarios por

[1] www.tfidf.com

[2] http://nlp.stanford.edu/software/tagger.shtml

medio de su interacción en la *web*. Mientras los usuarios sigan proporcionando información de su perfil en las redes sociales y naveguen por Internet, esto genera pruebas y evidencias para reconocerlos. Estos datos distribuidos del usuario pueden ser utilizados como una forma de obtener información para los sistemas que prestan servicios personalizados para sus usuarios o necesitan encontrar más información acerca de sus usuarios (Abel *et al.*, 2010).

Datos conectados de diferentes fuentes han sido utilizados para diferentes propósitos, como estandarización de *APIs*, por ejemplo `OpenSocial1` y personalización (Carmagnola y Cena, 2009).

Trabajos previos en detección de ciberacoso se han concentrado en el contenido de las conversaciones a pesar de que no se percataron de las características de los actores involucrados en el ciberacoso. Estudios sociales demostraron que un acosador masculino y femenino actúan de manera diferente.

Por ejemplo, las mujeres con estilos de comunicación agresiva, tienden a excluir a una persona de un grupo y empiezan a conspirar en contra de esa persona y los hombres tienden a usar más palabras y frases de indigación amenazadoras, según Chisholm (2006). Mientras que las mujeres menciona Argamon *et al.* (2003), utilizan más pronombres como <<yo, tú, ella, etc>> y los hombres manejan pronombres y artículos como <<un, una, el, eso>>.

Estos descubrimientos motivaron a la investigación de Dadvar *et al.* (2012), que menciona el efecto de las características basadas en el género para tratar el ciberacoso en redes sociales.

Nahar *et al.* (2012), propuso un metódo para detectar ciberacoso en redes sociales. Además presentaron un modelo de grafo para extraer una red de ciberacoso. Esto ha llevado a la identificación de los agresores más activos y víctimas a través de un algoritmo de clasificación.

Estos trabajos relacionados nos apoyaron en nuestra investigación para definir

el como detectar y extraer los mensajes agresivos que se encuentran en la red social.

3.3 Detección de ciberagresores

<<Un ciberagresor es una persona que hace uso de internet para cazar a víctimas con la intención de lastimarlas física, emocional, psicológica o económicamente. Los ciberagresores saben como manipular niños, creando una confianza basándose en una amistad que no existe >> (CyberSafety, 2015).

El centro Nacional de niños perdidos y explotados *(NCMEC)*, menciona que 1 de 7 jóvenes entre 10 y 17 años, tiene una experiencia de índole sexual a través del internet (NCMEC, 2008).

El reconocimiento de agresores es un tema que se encuentra dividido en dos áreas (Popescu y Grozea, 2012):

- Conocer la identidad de los agresores.
- Reconocer patrones para identificar a los agresores.

Para lograr conocer la identidad del usuario, una de las técnicas más utilizadas ha sido el filtrado de todas las conversaciones del agresor a través de una recopilación de frases clasificadas como *<<pervertidas>>* o con un valor especifico, por ejemplo, valores de peso TFIDF (Peersman *et al.*, 2012; Morris y Hirst, 2012; Parapar *et al.*, 2012; Eriksson y Karlgren, 2012).

El enfoque final es el de revisar los enunciados de las conversaciones previamente etiquetadas como posibles agresiones por medio del algoritmo propuesto por un método estandarizado para trabajar al nivel de enunciados (Kontostathis *et al.*, 2012; Hidalgo y Díaz, 2012; Kern *et al.*, 2012). Es por ello que, para lograr el reconocimiento de patrones y lograr identificar a los agresores, los agresores son detectados

en base a la obtención de su identidad, el cual se considera en la manera de entablar sus conversaciones (Popescu y Grozea, 2012).

Existen estudios que se enfocan en agresores sexuales a través de internet tales como los de Kontostathis (2009); McGhee *et al.* (2011) en estas investigaciones explican la teoría sobre técnicas de minería de texto que utilizan para distinguir entre conversaciones de agresores y víctimas, tal como se aplica una conversación uno a uno.

Investigaciones como la de Inches y Crestani (2012) no se basa solo en dos áreas, éste menciona que se puede identificar a los agresores en tres tareas: prefiltrado, extracción de características, clasificación.

Uno de los métodos más efectivos para la tarea del prefiltrado, de todas las conversaciones, fue realizado por Villatoro-Tello *et al.* (2012). Ellos despliegan algunos patrones específicos, por ejemplo, la existencia de un solo participante, ya sea con un mínimo de seis intervenciones por usuario o tres secuencias largas de caracteres sin reconocer.

Otros investigadores proponen tareas similares, aplicando un enfoque basado en reglas sobre diferentes características para diferentes métodos (Parapar *et al.*, 2012).

Para la segunda tarea, la extracción, se dividieron en dos grupos principales:

- Las características <<léxicas>>.
- Las características basadas en el <<comportamiento>>.

Las primeras investigaciones son aquellas que se puede derivar a partir del texto en bruto de la conversación, por ejemplo características con unigramas o bigramas como Villatoro-Tello *et al.* (2012); Morris y Hirst (2012); Eriksson y Karlgren (2012); Parapar *et al.* (2012), conteo de *emoticones* y conteo de valores de pesos TFIDF o similitud de cosenos.

Considerando las características dentro de una conversación, Vartapetiance y Gillam (2012); Hidalgo y Díaz (2012) utilizan el número de preguntas que realizan, la intención, captura la acción de los usuarios. Con esto se genera la creación de un conjunto sencillo de características para cada autor haciendo esto uno de los enfoques importantes. Este enfoque puede describir y desarrollar su agresor potencial.

Ayala *et al.* (2012); Kontostathis *et al.* (2012); Hidalgo y Díaz (2012); Peersman *et al.* (2012); Kern *et al.* (2012) aplican una estrategia de sumar los resultados de todas las líneas de la conversación para encontrar un único conjunto de características para cada autor.

Algunos investigadores no utilizan características en sí, si no que utilizan el modelo del lenguaje para dos participantes en el *chat* (Eriksson y Karlgren, 2012). Existen algunos enfoques que utilizan el modelo del lenguaje a un nivel de línea o de conversación. Otro ejemplo que utiliza Eriksson y Karlgren (2012) es el de reconocer el nombre de los participantes en las conversaciones (en sí mismo, con otros, en grupo).

Ševčíková y Šmahel (2009) proponen un método a través de entrevistas cara a cara con usuarios de internet para encontrar agresores y víctimas fundamentándose según su experiencia.

Se han propuesto diferentes perspectivas, para la clasificación de agresores y no agresores, como por ejemplo árboles de decisión (Kontostathis *et al.*, 2012), bosque aleatorio (Popescu y Grozea, 2012), así como con la teoría de clasificador bayesiano (Hidalgo y Díaz, 2012; Ayala *et al.*, 2012) y entropía máxima (Eriksson y Karlgren, 2012; Kern *et al.*, 2012).

Dentro de las diferentes perspectivas también se encuentran por medio de una comparación entre los clasificadores existentes, son las máquinas de soporte vectorial, las cuales son utilizadas por Morris y Hirst (2012); Parapar *et al.* (2012); Peersman *et al.* (2012); Villatoro-Tello *et al.* (2012). Y se han utilizado otros puntos de vista que han tenido mejor desempeño que las máquinas de soporte vectorial como por ejemplo

cuando aplican clasificadores con redes neuronales artificiales (Villatoro-Tello *et al.*, 2012).

Para poder identificar la manera de detectar a los agresores a través de la frecuencia de envío de mensajes agresivos que se encuentran en un caso de ciberacoso, nos apoyaron los trabajos relacionados en esta área.

3.4 Aplicaciones para la detección del ciberacoso

Con el crecimiento de ciberacoso entre niños y adolescentes, el tema importante que se espera de un adolescente es el discernir entre el bien y el mal. Es una responsabilidad de los padres proteger a sus hijos de los depredadores que se encuentran en el internet (K Jowalski R, 2010). En este tema se encuentran productos y redes comerciales que cumplen esta función como son *eBlaster* (Franklin, 2003), *Net Nanny* (Olah *et al.*, 2002) y *IamBigBrother* (Yang y Zhuang, 2015).

Packet sniffing es una de las alternativas más frecuentes para asegurar *chats*. *Packet sniffers* escanean todas las entradas y salidas de tráfico en la red y luego aplica un filtro para ver solamente los datos permitidos. Una de las razones importantes es que detectan a los depredadores basándose en una simple concordancia de palabras clave al igual que con una teoría de cancelar comunicaciones. Con ello brindando la exactitud de esta herramienta (Kontostathis *et al.*, 2010).

Otra aplicación que se encuentra en esta área es *Pidgin*, éste es un sistema de mensajería instantaneo. Trabaja con *Windows o ambiente Unix* y soporta múltiples protocolos como *AIM,MSN,IRC, y Yahoo*. A su vez, trabaja con `Facebook chat` utilizando *plugins* (Pidgin, 2015).

Una de las razones por las que es adecuado utilizar *Pidgin* es porque maneja un chat seguro. Pidgin se maneja en base a protocolos por lo que cuando aparece un

protocolo nuevo, la comunidad que elabora *Pidgin*, se asegura que no forme parte de un algoritmo de depredación.

SafeChat es un sistema que fue diseñado para trabajar con mensajería instantánea, forma parte de *Pidgin*, utiliza algoritmos que ayudan a clasificar si en el chat se encuentra participando depredadores potenciales (AOL, 2015).

Las aplicaciones que se han desarrollado para la detección automática de ciberacoso son muy reducidas (Dadvar y de Jong, 2012).

Entre estas aplicaciones se encuentra la detección de ciberacoso textual, realizada por Dinakar *et al.* (2011) en donde se utilizó una recopilación de comentarios realizados a videos de la página Youtube que trataban de temas sensibles relacionados a la raza, cultura, sexualidad e inteligencia.

Sus resultados muestran que la clasificación individual y binaria de temas sensibles puede superar la detección textual de ciberacoso comparada con el conjunto de datos que se obtuvieron. Ellos han ilustrado como sentido común del conocimiento aplicado en el diseño de software de una red social para la detección de ciberacoso.

Los autores tratan cada comentario por si mismo y no consideran otros aspectos del problema como sarcasmo en un diálogo. Llegaron a la conclusión de que teniendo en cuenta tales características será más útil en los sitios web de redes sociales y puede conducir a una buena solución del problema.

BullyTracer es un programa diseñado por Bayzick *et al.* (2011) para detectar presencia de diferentes tipos de ciberacoso en una conversación dentro de un foro.

Esta aplicación analiza todos los archivos de un directorio proporcionado utilizando reglas basadas en algoritmos.

El uso de n-gramas, *skip grams*, y palabras especiales como “yo” se han utilizado para construir un modelo vectorial, esta aplicación se ha desarrollado para detectar comentarios que son considerados insultos hacia participantes de un blog/foro

de conversación (Goyal y Kalra, 2013).

Esta área de aplicaciones para la detección de ciberacoso apoyo a nuestra metodología propuesta a generar un análisis de lo que se encuentra en el mercado, las aplicaciones tecnológicas que existen y el como poder lograr una detección de ciberacoso utilizando un recurso en línea.

3.5 Áreas relacionadas con la detección automática del ciberacoso

Nahar *et al.* (2012) menciona que los desafíos en la lucha contra el ciberacoso son:

1. Detectar el acoso en línea cuando está ocurriendo.
2. Reportar a las autoridades.
3. Reportar a los servicios de internet.
4. Las escuelas, deben de tener el propósito de prevenir y tomar conciencia en los adolescentes
5. Identificar al agresor y a sus víctimas.

Por lo que se han relacionado varios temas en cuanto a la detección del ciberacoso (Dadvar y de Jong, 2012); investigaciones basadas en paradigmas de minería de texto y análisis de sentimiento (Pang y Lee, 2008), detección de mal comportamiento en los usuarios que utilizan chats (Villatoro-Tello *et al.*, 2012), detección de spam (Tan *et al.*, 2010), detección de depredadores sexuales en línea (McGhee *et al.*, 2011) y detección de ciberterrorismo (Simanjuntak *et al.*, 2010).

Otra tema que se enfoca en la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas que un usuario busca y

puede aportar en la detección del acoso es la recuperación de información (Tolosa y Bordignon, 2008).

El área que recibe atención en el tema de ciberacoso es la de encontrar depravados sexuales *" sexual harassment "*, en donde se han realizado avances destacados como el trabajo de Potha *et al.* (2016) que utilizando herramientas computaciones y utilizando metodologías que se encuentran dentro del área de las ciencias biológicas se encontraron patrones que apoyan a detectar a los integrantes de estos casos de acoso.

Sin embargo, se le ha prestado muy poca atención a la detección del ciberacoso de manera automática. La detección de ciberacoso y el desarrollo de medidas preventivas subsecuentes son las principales líneas en la lucha contra el ciberacoso (Nahar *et al.*, 2012).

Con estas investigaciones nos percatamos que para nuestra investigación es conveniente tratar de encontrar una metodología que propusiera técnicas para encontrar el ciberacoso de manera automática, ya que como se menciona es una área que falta por desarrollarse.

En la tabla 3.4 se presenta una comparación de los trabajos que se han mencionado en este cápitulo junto con la solución propuesta en esta tesis.

Tabla 3.1: Comparación de trabajos relacionados

Trabajos \ Características	Metodología	Detectar mensajes agresivos	Detectar agresores	Detectar frecuencia de envío de mensajes agresivos	Uso de Redes Sociales	Apoyar a detectar agresividad en linea	Uso de nivel de agresividad	Uso de sistemas difusos	Detectar tipos de ciberacoso	Detectar perfiles de usuario
Salmivalli *et al.* (1996)	Social	-	-	-	-	-	-	-	✓	-
Chen *et al.* (2012)	Social	-	-	-	-	-	-	-	-	✓
Walrave y Heirman (2011)	Social	-	-	-	-	-	-	-	✓	-
Nahar *et al.* (2012)	Social	✓	✓	-	-	-	-	-	-	✓
Del Rey *et al.* (2015)	Social	-	-	-	✓	-	-	-	-	✓
Patchin y Hinduja (2006)	Social	-	-	-	-	-	-	-	✓	-
Willard (2007)	Social	-	-	-	-	-	-	-	✓	-
Maher *et al.* (2008)	Social	-	-	-	-	-	-	-	✓	-
Nadali *et al.* (2013)	Tecnológica	-	-	-	-	-	-	-	✓	-
Yin *et al.* (2009)	Tecnológica	-	✓	✓	✓	-	-	-	-	-
Al-garadi *et al.* (2016)	Tecnológica	✓	-	-	✓	-	-	-	✓	-
Chisholm (2006)	Tecnológica	-	-	-	-	-	-	-	✓	✓
Dinakar *et al.* (2011)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Reynolds *et al.* (2011)	Tecnológica	✓	-	-	✓	✓	-	-	✓	-
Abel *et al.* (2010)	Tecnológica	-	-	-	-	-	-	-	-	✓
Carmagnola y Cena (2009)	Social	-	-	-	-	-	-	-	-	✓

Tabla 3.2: Comparación de trabajos relacionados

Características / Trabajos	Metodología	Detectar mensajes agresivos	Detectar agresores	Detectar frecuencia de envío de mensajes agresivos	Uso de Redes Sociales	Apoyar a detectar agresividad en linea	Uso de nivel de agresividad	Uso de sistemas difusos	Detectar tipos de ciberacoso	Detectar perfiles de usuario
Chisholm (2006)	Tecnológica	-	-	-	-	-	-	-	✓	✓
Argamon *et al.* (2003)	Social	-	-	-	-	-	-	-	-	✓
Dadvar *et al.* (2012)	Tecnológica	-	-	-	-	-	-	-	✓	-
CyberSafety (2015)	Tecnológica	-	✓	-	-	-	-	-	-	-
NCMEC (2008)	Social	-	-	-	-	-	-	-	-	✓
Popescu y Grozea (2012)	Tecnológica	-	✓	-	-	-	-	-	-	-
Peersman *et al.* (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Morris y Hirst (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Parapar *et al.* (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Eriksson y Karlgren (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Kontostathis *et al.* (2012)	Tecnológica	✓	-	-	-	-	-	-	-	✓
Hidalgo y Díaz (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Kern *et al.* (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Kontostathis (2009)	Tecnológica	-	✓	-	-	-	-	-	-	✓
McGhee *et al.* (2011)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Inches y Crestani (2012)	Tecnológica	-	✓	-	-	-	-	-	-	✓

Tabla 3.3: Comparación de trabajos relacionados

Características / Trabajos	Metodología	Detectar mensajes agresivos	Detectar agresores	Detectar frecuencia de envío de mensajes agresivos	Uso de Redes Sociales	Apoyar a detectar agresividad en linea	Uso de nivel de agresividad	Uso de sistemas difusos	Detectar tipos de ciberacoso	Detectar perfiles de usuario
Villatoro-Tello *et al.* (2012)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Vartapetiance y Gillam (2012)	Tecnológica	-	✓	-	-	-	-	-	-	✓
Ayala *et al.* (2012)	Tecnológica	✓	-	-	✓	-	-	-	-	✓
Ševčíková y Šmahel (2009)	Tecnológica	-	✓	-	-	-	-	-	-	✓
K Jowalski R (2010)	Tecnológica	-	-	-	-	-	-	-	✓	-
Franklin (2003)	Tecnológica	-	-	-	-	-	-	-	✓	-
Olah *et al.* (2002)	Tecnológica	-	-	-	-	-	-	-	✓	-
Yang y Zhuang (2015)	Tecnológica	-	-	-	-	-	-	-	✓	-
Kontostathis *et al.* (2010)	Tecnológica	✓	-	-	-	-	-	-	✓	-
Pidgin (2015)	Tecnológica	✓	-	-	-	-	-	-	-	-
AOL (2015)	Tecnológica	✓	-	-	-	-	-	-	-	✓
Dadvar y de Jong (2012)	Tecnológica	-	-	-	-	-	-	-	✓	-
Dinakar *et al.* (2011)	Tecnológica	✓	-	-	-	-	-	-	✓	-
Bayzick *et al.* (2011)	Tecnológica	-	-	-	-	-	-	-	✓	-
Goyal y Kalra (2013)	Tecnológica	✓	✓	-	-	-	-	-	-	-

Tabla 3.4: Comparación de trabajos relacionados

Características / Trabajos	Metodología	Detectar mensajes agresivos	Detectar agresores	Detectar frecuencia de envío de mensajes agresivos	Uso de Redes Sociales	Apoyar a detectar agresividad en linea	Uso de nivel de agresividad	Uso de sistemas difusos	Detectar tipos de ciberacoso	Detectar perfiles de usuario
Pang y Lee (2008)	Tecnológica	✓	✓	-	-	-	-	-	✓	-
Tan *et al.* (2010)	Tecnológica	✓	-	✓	-	-	-	-	-	-
McGhee *et al.* (2011)	Tecnológica	✓	✓	-	-	-	-	-	-	-
Simanjuntak *et al.* (2010)	Tecnológica	✓	✓	-	-	-	-	-	✓	-
Tolosa y Bordignon (2008)	Tecnológica	✓	-	-	-	-	-	-	✓	✓
Potha *et al.* (2016)	Tecnológica	-	-	-	✓	-	-	-	✓	✓
Solución propuesta	Tecnológica	✓	✓	✓	✓	✓	✓	✓	✓	✓

3.6 Resumen

En este cápitulo se describen los trabajos e investigaciones realizados en el estado del arte que ayudan a detectar el ciberacoso. El ciberacoso ha sido estudiado en el ámbito social; y de cómo por medio de metodologías, apoyan a su detección y con ellas obtener pruebas para llegar a levantar una denuncia a las autoridades pertinentes.

En el estado del arte se encuentran investigaciones que apoyan a detectar el ciberacoso, por medio de aplicaciones, también de como se puede detectar a los agresores que generan el ciberacoso. Sin embargo los trabajos donde apoyen a detectar el ciberacoso en línea y de manera automática son pocos los que se han desarrollado por completo, para llegar a esta finalidad.

Capítulo 4

Metodología

Si no conozco una cosa, la investigaré.

Luis Pasteur

Este trabajo de investigación presenta una metodología enfocada en la identificación de casos de ciberacoso dentro de una red social, de la cual se extraen mensajes de texto calificados como agresivos, sus emisores agresores y las víctimas de los mismos. Esta metodología incluye etapas sucesivas.

A raíz de la definición de ciberacoso considerada como <<un acto **agresivo intencional** realizado por un grupo de individuos o un individuo, utilizando vías electrónicas para contactar de **manera consecutiva** a una **víctima**>> (Smith, 1999), se proponen las etapas para esta metodología, ver figura 4.1:

1. Detección de mensajes de texto agresivos.
2. Detección de posibles agresores y posibles víctimas.
3. Detección de casos de ciberacoso (agresor, víctima y mensajes).

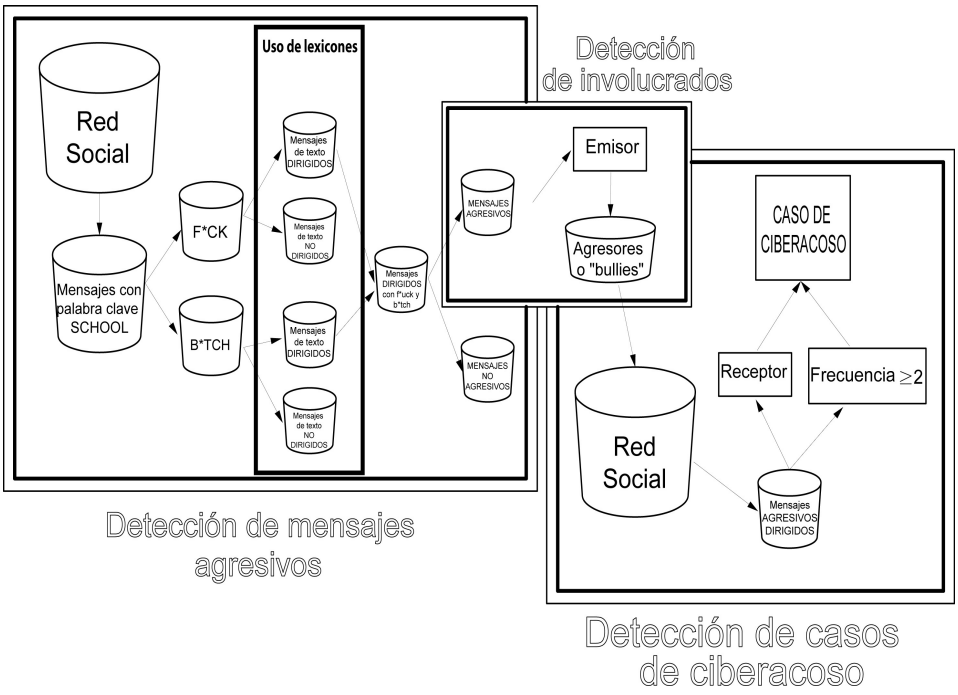

Figura 4.1: Proceso para detectar casos de ciberacoso en una red social.

Ya que el ciberacoso es considerado como un acto agresivo intencional, se propone como primer etapa de esta metodología la detección de agresividad en los mensajes de texto dentro de una red social. Se considera una red social como una estructura que se encuentra compuesta por individuos los cuales manejan una relación cibernética y se pueden enviar mensajes personales entre ellos (Nahar *et al.*, 2012), ver figura 4.2.

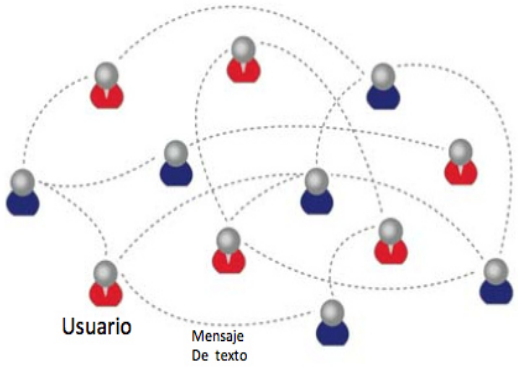

Figura 4.2: Ejemplo de lo que se considera como una red social.

En la definición de ciberacoso se especifica que *es intencional y realizado por un grupo de individuos o un individuo*. Por esta razón, la segunda etapa consiste en la detección de posibles agresores o *bullies* y sus posibles agredidos o víctimas. Una vez determinados cuáles son los mensajes de texto agresivos, nos enfocamos en quiénes son los emisores y receptores que envían estos mensajes.

La última parte de la definición de ciberacoso menciona que *se contacta de manera frecuente, a una víctima o víctimas*. Así que la tercer etapa del estudio considera la frecuencia del envío de mensajes de texto agresivos entre los posibles agresores y posibles víctimas detectados, facilitando la detección de casos de ciberacoso.

El capítulo se organiza de la siguiente manera: Primero se presenta cómo se detectan los mensajes de texto agresivos, en donde se explica sobre como se obtuvieron los mensajes de texto de la red social y cómo se detecta su nivel de agresividad. Se continúa con la detección de los involucrados en el envío de estos mensajes de texto. Y por último, se describe la detección de casos de ciberacoso, en donde se detalla la metodología de cómo se logra detectar estos casos.

4.1 Detección de mensajes de texto agresivo

Hay una cantidad considerable de maneras de hacer ciberacoso, ver imagen 4.3 (Hinduja y Patchin, 2008). El ciberacoso se puede generar a través de diferentes medios como: videos, imágenes y correos electrónicos. En general, el uso de mensajes de texto instantáneos es lo más común en el acoso cibernético, según un estudio de Kowalski *et al.* (2014). Por esta razón, nos enfocamos en los mensajes de texto, dejando para trabajo a futuro los otros medios para generar ciberacoso.

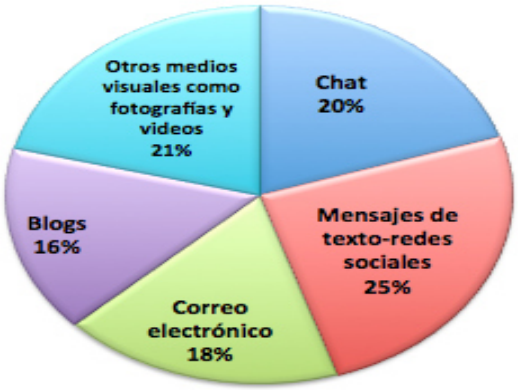

Figura 4.3: Porcentajes de uso de medios electrónicos

El uso de redes sociales en Internet refleja una tendencia a la alta en ámbitos escolares, laborales y particularmente en la vida social. La interacción, en estas redes sociales, permite que Internet sea una plataforma donde se genere este tipo de agresividad (Kowalski *et al.*, 2014).

Es por esto que la detección de agresividad y su frecuencia es una tarea importante en este proceso de búsqueda de ciberacoso en redes sociales. Esto debido a que se ha vuelto una práctica muy común el mandar o *postear* mensajes para expresar diferentes sentimientos, incluida la agresión, entre los individuos (Dinakar *et al.*, 2011).

Lo anterior permite considerar de importancia para este estudio las siguientes

tareas:

- Obtención de mensajes de texto posiblemente agresivos.
- Asignación de nivel de agresividad a cada mensaje de texto.

4.1.1 Obtención de mensajes de texto posiblemente agresivos

Una red social es una red masiva. Se conoce como redes masivas a aquellas redes donde cualquier persona puede participar, con el fin de compartir mensajes de texto, imágenes, videos y de hacer nuevos contactos (Sánchez, 2013). Por ejemplo, la red social *Twitter* cuenta con más de 100 millones de mensajes de texto diarios (Sánchez, 2013). Es por esto que para la investigación es necesario hacer una búsqueda para obtener los mensajes de texto requeridos, que clasifiquen como mensajes agresivos.

Los mensajes considerados agresivos son los que tienen la intención de herir, denigrar, molestar o insultar a otra persona u otras personas (Sameer y Patchin, 2008). Estos mismos mensajes generan reacciones en quienes los reciben y material de estudio para otras áreas de investigación, como en las áreas humanistas, entre ellas psicología.

Para lograr identificar este tipo de mensajes en una red social se realiza lo siguiente:

1. Filtrar los mensajes de texto de la red social utilizando palabras clave en donde es posible que existan casos de ciberacoso. Para esta investigación se consideran los mensajes de texto con la palabra clave: *school* [1]. Se utiliza esta palabra en específico, porque se pretende buscar agresividad en los textos y este tema es propicio para el desarrollo de esta búsqueda.

[1] Se utiliza el idioma inglés, más adelante se justifica el porque se toma esta decisión.

2. De los mensajes de texto filtrados, se realiza una búsqueda para encontrar los que incluyan palabras agresivas, considerando como muestra para esta investigación utilizamos las palabras *f*ck* y *b*tch* (Dadvar *et al.*, 2012; Ptaszynski *et al.*, 2010).

3. Al tener los mensajes filtrados por las palabras agresivas, se escogen los mensajes de texto dirigidos o direccionados. Debido a que se ha detectado que la participación identificada como negativa de un miembro o posible miembro agresivo de la red es relativa a su contenido dirigido hacia otro usuario (Bhutani *et al.*, 2012). Es decir, que si el usuario escribe de manera agresiva, es posible que esta agresividad se vea plasmada hacia un receptor en específico.

Cuando un mensaje de texto va dirigido hacia una persona o a cierto grupo de personas y también cuenta con malas palabras o palabras ofensivas puede tratarse de ciberacoso. La combinación de agresión y personalización del mensaje de texto puede llegar a ser muy hiriente para el usuario al que se le dirige el mensaje de texto (Dinakar *et al.*, 2011).

En *Twitter*, por ejemplo, un mensaje de texto direccionado o dirigido, es un mensaje que normalmente contiene la arroba *"@"* seguido del nombre de usuario. Por su parte *Facebook*, identifica con color azul el nombre del usuario a quien va dirigido, como se puede ver en la tabla 4.1.

En el caso de los nombres de usuario que cuentan con palabras agresivas, como por ejemplo, *"@m*therf*ck*er"* no se consideran como parte del texto analizado durante el procesamiento del sistema ya que se considera en el filtrado del mismo.

En un mensaje dirigido o direccionado es adecuado, para esta etapa del proceso, considerar la emoción, en este caso el de la agresión; pueden existir consecuencias negativas al prestar atención en las emociones que se encuentran en el contexto de los mensajes e inclusive en el comportamiento humano (Breazeal, 2003; Liu *et al.*, 2003). El no reconocer emociones sin reconocer el texto en donde se expresan no

Tabla 4.1: Ejemplos de mensajes de texto direccionados y no direccionados en redes sociales

	Twitter	Facebook
Texto direccionado	"@Abzzurd He's better though "	"Mario Carrillo i love you "
Texto no direccionado	"My life is sooooo damn good"	"I love fridays!!"

sería del todo suficiente para realizar aplicaciones en el mundo real (Ptaszynski *et al.*, 2010). Con aplicaciones en el mundo real se refiere a metodologías que apoyen a detectar varios problemas como ciberacoso, pornografía y otro tipo de incidencias que afecten a la sociedad.

4.1.2 Asignación de nivel de agresividad a cada mensaje de texto

Una vez teniendo un conjunto de datos con los posibles mensajes de texto agresivos, al cual llamaremos $<<M1>>$, se procede a asignar el nivel de agresividad para cada uno de estos mensajes. Este nivel se establece en una escala de cero a diez, en donde cero significa *nada agresivo* y diez significa *muy agresivo*. Tomamos en cuenta esta escala porque se considera apropiada para este contexto ya que no es muy extensa ni muy reducida, en comparación con una escala mayor como de cero a cien.

El nivel de agresividad de cada mensaje que conforma *M1* se crea por medio de un conteo de palabras agresivas encontradas en relación a la cantidad de palabras que conforman el mensaje, considerando palabras agresivas las encontradas en un *lexicón*. El lexicón es una lista de palabras. Esta lista se obtuvo de la página `noswearing.com`; esta página contiene una gran cantidad de malas palabras, las cuales se encuentran en el idioma inglés (Noswearing.com). Se utiliza esta lista debido a que obtuvo buenos resultados en nuestros experimentos realizados para la detección de agresividad en los mensajes de texto.

El estudio del ciberacoso, es multidisciplinario e incipiente en el área de tecnologías de información. Debido a esto, la información de apoyo para esta investigación se encuentra en su mayoría en el idioma inglés en comparación con otros idiomas como el español. Por lo que la información en inglés se encuentra accesible para este estudio.

El proceso para detallar cómo se genera el nivel de agresividad sc_i de cada mensaje de texto que se encuentra en *M1*, es el siguiente:

1. Cada mensaje de texto de *M1* se divide en palabras. Se realiza la división de palabras para poder comparar cada palabra del mensaje con cada una de las palabras que conforman el lexicón.

2. Al estar realizando la comparación de cada palabra del mensaje de texto con las de lexicón, se utiliza la ayuda de un contador. El contador va a ir sumando de uno en uno, la cantidad de veces que encuentra una de las palabras del lexicón, en cada mensaje. Este dato lo llamaremos *contador palabras*.

3. Con ayuda de un contador adicional, se considera el numero total de palabras que conforman el mensaje, a este dato lo llamaremos *total*.

4. Se divide el *contador palabras* y *total* generando un *cociente* para cada uno de los mensajes. Este *cociente* se va convirtiendo en un *cociente máximo*.

5. el *cociente máximo* se convierte en máximo cuando el *cociente máximo* sea menor que el *cociente*.

6. Se realiza una división entre cada *cociente* de los mensajes y el *cociente máximo*, al terminar se multiplica por diez, para generar el valor entre cero y diez, los valores que conforman la escala.

7. Estos valores generados son los niveles de agresividad para cada uno de los mensajes que conforman *M1*.

Una manera de explicar matemáticamente, lo que se acaba de mencionar, sobre como se origina este nivel de agresividad utilizando el lexicón, es obteniendo la frecuencia de palabras ofensivas f_i que se encuentran en $M1$, normalizando esta frecuencia utilizando el valor de *cociente máximo* que se encuentra en *M1* . La frecuencia f_i de palabras ofensivas en cada mensaje de $M1$ se genera calculando la proporción de malas palabras $M1$, de tal manera que

$$f_i = \frac{o_i}{n_i}, \tag{4.1}$$

donde o_i y n_i son, respectivamente, el total de palabras agresivas y el total de palabras (ambos de $M1$).

El valor sc_i es finalmente calculado, normalizando la frecuencia relativa con el *cociente máximo* $f_{\max}$ encontrado y multiplicando el resultado por diez, generando el valor tal como se encuentra nuestra escala $[0, 10]$:

$$sc_i = (10)\left(\frac{f_i}{f_{\max}}\right). \tag{4.2}$$

Por ejemplo, asumiendo que w_1 y w_2 son palabras agresivas en un documento $M1 = \{w_1, w_2, w_3, w_4\}$. En este caso, $f_i = \frac{2}{4} = 0.5$; si $f_{\max} = 0.6$, entonces
sc_i =$(10)(\frac{0.5}{0.6}) = 8.3$.

No utilizamos los métodos supervisados porque los conjuntos de entrenamiento tienen el problema de las clases *desbalanceadas*, es decir, la mayoría de los mensajes son no agresivos y relativamente pocos son agresivos o muy agresivos. Esto nos llevaría a tener que buscar métodos de muestreo y adicionalmente métodos de ensamble, por lo que se deja esto para trabajo futuro.

En esta etapa obtenemos un número, este representa el nivel de agresividad para cada uno de los mensajes de texto obtenidos de la red social. Con estos datos se genera un conjunto de posibles mensajes de texto agresivos. Al contar con este conjunto de mensajes, se prosigue a la siguiente etapa del proceso, la detección de

los posibles agresores y las posibles víctimas.

4.2 Detección de los involucrados

Un usuario que presenta una actividad constante en la red social y se le detecte que sus mensajes cuentan con agresividad se considera un *agresor o acosador*.

A un usuario se le va a considerar un *agresor o acosador* si los mensajes de texto que envía cuentan con un nivel de agresividad igual o mayor a un nivel de cinco. A un usuario se le va a considerar *agredido o víctima* si los mensajes de texto que recibe cuentan con un nivel de agresividad igual o mayor a un nivel de cinco. Esta calificación de cinco para el nivel de agresividad se selecciona debido a que es la media de nuestra escala de agresividad.

El proceso para determinar si un emisor es un agresor y un receptor es víctima del agresor, es el siguiente:

1. Se ordenan los mensajes de *M1*, según su nivel de agresividad, de mayor a menor.
2. Los emisores de los mensajes que cuenten con nivel de agresividad igual o mayor a cinco, son considerados posibles agresores o *bullies*.
3. Con los posibles agresores, se genera un subconjunto definido como *semillero*.
4. En la red social se realiza una búsqueda de las conversaciones de cada uno de los posibles agresores, para revisar sus comentarios y detectar si los mensajes que envía, se consideran agresivos. Formando otro conjunto de mensajes de texto, pero estos solo pertenecen a los posibles agresores detectados. Este conjunto de datos lo definiremos como S_1.
5. Se obtienen conversaciones, de las cuales se consigue la frecuencia con la que se envían mensajes agresivos entre los involucrados detectados; si el mensaje

es considerado agresivo el emisor es un agresor o *bully*.

6. Por ende los receptores de estos mensajes, es decir, las personas a quienes se le dirigen estos mensajes agresivos, son consideradas víctimas.

Un usuario es un agresor o una víctima basándose en los mensajes detectados como agresivos que envía o recibe respectivamente. Por eso, un usuario es asignado como un agresor y se le asigna su víctima según su nivel de agresividad que se encuentren en sus conversaciones.

4.3 Detección de casos de ciberacoso

En la primer etapa se trabajó con mensajes individuales o aislados. En la segunda etapa se identificaron los emisores de estos mensajes. Por último, en la tercera etapa se detectarán los posibles casos de ciberacoso.

En esta tercera etapa se procede a registrar, por un periodo de tiempo, las conversaciones que tienen los agresores con sus posibles víctimas. Se rastrea la frecuencia con la que estos involucrados se comunican de manera agresiva; esto se realiza considerando la última parte de la definición de ciberacoso es: << *un acto agresivo, intencional realizado por un grupo de individuos o un individuo, utilizando medios electrónicos de comunicación como: Messenger, Facebook, Twitter, Youtube, etc; para contactar de manera frecuente, por un período de tiempo, a una víctima o víctima*>>, como se ejemplifica en la figura 4.4.

Se desarrolla, como primer paso, la detección de las conversaciones del usuario que se detecta como agresor, en un periodo considerado de tiempo.

El periodo de tiempo considerado para esta investigación es de seis meses, ya que es la cantidad de tiempo más lejana que nos permite trabajar en la red social, de donde obtenemos la información, además es lo que permiten manejar dentro de un promedio del historial de cada usuario.

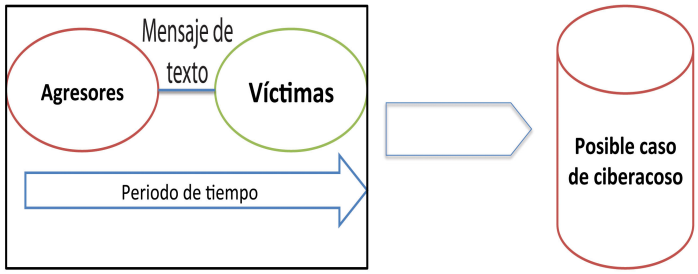

Figura 4.4: Ejemplo de factores necesarios para detectar un posible caso de ciberacoso.

De estas conversaciones se obtiene: el mensaje enviado (m), el o los receptores (v) y la fecha en la que se emitió cada uno de estos mensajes ($frec$). Teniendo estos datos se forma una base de datos en donde se procede a generar su respectivo nivel de agresividad sc_i y ordenarlos de mayor a menor (M_2). M_2 es el conjunto de conversaciones que son consideradas agresivas.

Para detectar el valor de nivel de agresividad de las conversaciones se genera un promedio de los mensajes que conforman las conversaciones ($conv$); es decir, si tenemos $conv_1 = \{m_1, m_2, m_3\}$ se generan sus niveles de agresividad correspondientes como se mencionó en la sección de detección de mensajes de texto agresivos.

Supongamos que los niveles que le corresponden a cada uno de los mensajes que se encuentran en $conv_1$ son: $sc_i1 = 8.0$, $sc_i2 = 5.0$ y $sc_i3 = 3.5$, respectivamente.

El promedio para $conv_1$ sería $\frac{(8.0+5.0+3.5)}{3} = 5.5$, este valor de nivel de agresividad es mayor de 5, por lo que esta conversación es considerada agresiva.

Al tener identificado los niveles de agresividad por conversación con su agresor y sus respectivas víctimas, se realiza un promedio total con estas conversaciones $prom_{conv}$.

Este promedio se genera sumando los niveles de agresividad de cada conversacion que conforma M_2 y dividiéndolo entre la cantidad de conversaciones.

Por ejemplo, supongamos que tenemos $M_2 = \{conv_1, conv_2, conv_3\}$, y sus niveles de agresividad correspondientes son 5.5, 4.83 y 11.52, entonces el promedio para generar $prom_{conv}$ sería $\frac{(5.5+4.83+11.52)}{3} = 7.28$,

Con los promedios generados para cada conversación, por usuario agresor y obteniendo el *promedio total* ($prom_{conv}$) se comparan los resultados de todos los promedios de las conversaciones que pertenecen al mismo agresor. Si el promedio de la conversación *conv* es mayor al *promedio total* ($prom_{conv}$) , esta conversación es un caso de ciberacoso detectado. En nuestro ejemplo el $prom_{conv}$ tiene un valor de 7.28 por lo que la conversación que sería considerada ciberacoso es $conv_3$.

4.4 Resumen

Con este trabajo se propone una metodología para encontrar casos de ciberacoso en una red social. Esta metodología se encuentra compuesta por tres etapas, ver figura 4.1. Cada una de las etapas surge en base de la definición de ciberacoso.

En la primer etapa que corresponde a la detección de mensajes de texto agresivos se selecciona la red social con la que se va a trabajar. De esta red social se seleccionan los mensajes de texto por medio de palabras clave, para buscar agresividad. Se genera y utiliza el nivel de agresividad que sirve para conocer qué tan agresivo es el mensaje de texto, en una escala entre cero (nada agresivo) y diez (muy agresivo).

Al contar con los mensajes de texto agresivos se continúa con la segunda etapa del proceso. En esta etapa se detectan quiénes son los involucrados que envían estos mensajes. Con esto se logra identificar quién puede ser el o los posibles agresores o *bullies* y quién puede ser el o los agredidos o víctimas.

Al obtener a los posibles agresores, se prosigue con la tercera etapa que es detectar los casos de ciberacoso. Para esto se requiere identificar las conversaciones

de los mensajes emitidos por estos agresores en un periodo de tiempo determinado. Para esta investigación se utilizaron seís meses.

Contando con las conversaciones del agresor en un periodo de tiempo determinado, se calcula el nivel de agresividad para cada uno de estas conversaciones. Con el nivel de agresividad se genera un promedio total entre los niveles de agresividad de cada conversación y el total de las conversaciones que se encuentran en la base de datos. Las conversaciones que sean mayores al promedio total de nivel de agresividad generado se consideran casos de ciberacoso.

En el siguiente capítulo << Experimentos y resultados >> se presentan los experimentos realizados en esta investigación. Además se muestran los resultados que se obtuvieron de estos experimentos basados en la metodología que se acaba de explicar en este capítulo.

Capítulo 5

Experimentos y Resultados

La formulación de un problema, es más importante que la solución.

- Albert Einstein.

En este capítulo se presentan los experimentos que se realizaron para comprobar si es posible detectar ciberacoso en las redes sociales a través de herramientas que se encuentran dentro del área de análisis de sentimiento y de minería de texto, siendo esto la hipótesis planteada. Este capítulo se divide en tres secciones, que son: caso de estudio, detección de mensajes de texto agresivos y detección de casos de ciberacoso.

5.1 Introducción

Para llevar a cabo la identificación automática del ciberacoso en redes sociales, se requiere conocer los mensajes de texto, quiénes envían estos mensajes y el tiempo definido durante el cual se realizan estos mensajes. Esto en base a la definición de ciberacoso que mencionamos en la metodología.

Sabemos que para poder alcanzar a identificar el ciberacoso una de las etapas

importantes es la detección de mensajes de texto agresivo. Es por ello que definimos una escala de agresividad (cero a diez, donde cero es nada agresivo y diez es muy agresivo).

Utilizamos además diferentes procedimientos para evaluar los mensajes de texto en términos de agresión; dentro de estos procedimientos se incluye una comparación con los resultados provenientes de trabajar con lexicones y lógica difusa.

Para realizar las evaluaciones de estos enfoques, se extrajeron dos conjuntos de datos de la red social `Twitter` (nuestro caso de estudio), ver tabla 5.2, los cuales se encuentran en una ventana de tiempo de un rango de 6 meses (Febrero a Julio del 2015), como se menciona anteriormente en la metodología. Comparamos los valores producidos de manera automática con los valores generados por un grupo de evaluadores. Nuestros resultados, en general, muestran que varios de los enfoques son factibles, en específico aquellos que combinan características diferentes.

En nuestra investigación, además de realizar experimentos para la detección de texto agresivo, se realizaron experimentos para la detección de ciberacoso. Estos experimentos se basan en la detección de agresividad en las conversaciones de los agresores con sus víctimas, en un periodo de tiempo.

También se realizaron evaluaciones para corroborar que la agresión detectada por nuestra metodología es considerada de igual manera por los evaluadores. Con ello comprobamos que nuestra hipótesis es aceptada.

El capítulo se organiza de la siguiente manera: Primero se presenta el caso de estudio, en donde se explica sobre cómo se obtuvieron los conjuntos de datos obtenidos de la red social `Twitter`. Se continúa con la explicación de los experimentos que se realizaron para lograr la detección de agresividad en los mensajes de texto. Y por último, se describe la detección de casos de ciberacoso, en donde se detallan las pruebas realizadas para obtener los resultados de esta investigación.

5.2 Caso de estudio

La investigación se centra en detectar casos de ciberacoso en una red social. Así que al comenzar nuestra investigación, encontramos que las dos redes sociales, actualmente, más populares son Facebook y Twitter. Estas dos redes son populares tanto en visitas como en contenido de ciberacoso, como se muestra en la figura 5.1 y en la tabla 5.1. Los comentarios que se encuentran en estas redes sociales suelen provocar un impacto importante tanto positivo como negativo en los usuarios que las utilizan (Chen *et al.*, 2012).

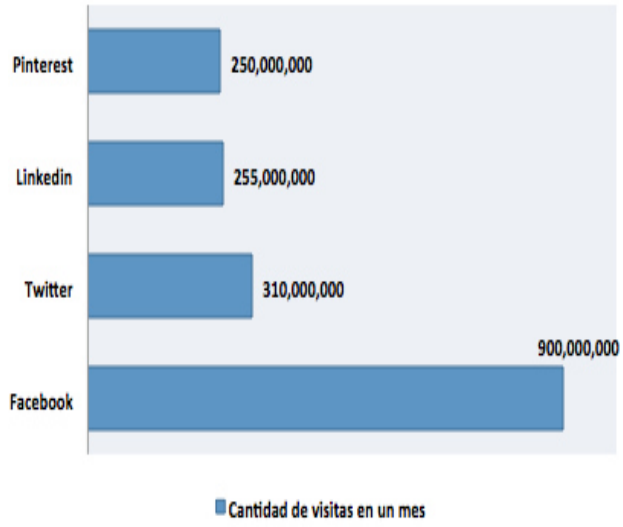

Figura 5.1: Redes sociales más populares (Bussines guide Inc, 2015)

El conjunto de mensajes de texto que contempla nuestra investigación, fue obtenida de la red social `Twitter`, ya que es una de las redes sociales más populares. `Twitter (http://www.twitter.com)`, se basa en la propagación de pequeños comentarios llamados *tuits*, los cuales son publicados por los usuarios y distribuidos automáticamente a otros usuarios que deseen recibir actualizaciones de estos. Cada comentario está limitado a 140 caracteres de longitud y diariamente se publican alrededor de 200 millones (Sui, 2015), los cuales pueden ser vistos por cualquier usuario

Tabla 5.1: Casos de mensajes agresivos que se encontraron en una base de datos perteneciente a redes sociales (Sameer, 2015)

Red Social	Casos en un mes
Facebook	92.6 %
Twitter	23.8 %
MySpace	17.7 %
Instant Messenger	15.2 %

registrado. Estas características lo hacen interesante para el análisis de sentimiento debido a que ofrece un conjunto de funciones o métodos llamados *API* (Interfaz de Programación de Aplicaciones) que permiten la comunicación de una computadora con la red social. Una *API* es una interfaz de comunicación entre dos componentes de software que facilita la labor de comunicación y programación, por lo que los mensajes pueden ser procesados rápidamente por los algoritmos y es relativamente fácil obtener una cantidad suficiente de comentarios para un análisis. Utilizando la *API* de `Twitter` nos fue posible crear un programa para obtener 111,381 mensajes de texto.

Tabla 5.2: Cantidad de mensajes de texto que se obtuvieron de la *Twitter API*

Clasificación	Mensajes de texto
Base de datos completa	111,381
Direccionados	12,705
Filtrados con la palabra ofensiva f**k	281
Filtrados con la palabra ofensiva b**ch	110

`Twitter` otorga la autorización para poder obtener la información que se requiere para fines de investigación, a diferencia de otras redes sociales como Facebook, que en el momento que se empezó con esta investigación obtener datos de Facebook era muy complicado, además que no se conseguía una cantidad considerada de in-

formación debido a los permisos que esta red social maneja. La información que se utiliza en este conjunto de datos cuenta con la autorización de los usuarios debido a que la aplicación considera las configuraciones de privacidad de cada una de las cuentas de donde se obtuvo la información.

Los registros que se utilizaron para el desarrollo de esta investigación se obtuvieron de manera legítima y sin obstruir la privacidad de los usuarios que colocaron estos mensajes en su perfil. Esto debido a que se realizan basándose en las reglas de desarrollo de la aplicación de `Twitter`.

Al contar con la red social elegida, se procedió a seleccionar los mensajes de texto donde posiblemente existe agresividad o ciberacoso para la investigación, tal como se explicó en la metodología en el punto 4.1.1 y como se puede ver en la tabla 5.2.

Se requiere buscar mensajes de texto agresivos, debido a que nuestra definición de ciberacoso, menciona en sus primeras líneas que es un acto agresivo intencional, es por ello que se requiere buscar agresividad en los mensajes de texto como primer paso para identificar si puede existir o no un caso de ciberacoso.

Para lograr la detección de los mensajes de texto agresivos, se requiere seleccionar primero la red social que se va a utilizar. Además para este trabajo de investigación, hay que definir que el objetivo de nuestros experimentos se maneja en dos sentidos. Por un lado, queremos comparar diferentes enfoques propuestos, y por otro lado, también se desea tener una noción de cual de estos enfoques cuenta con un mejor soporte para la detección de texto agresivo. Se evaluaron diferentes enfoques para escoger el enfoque que nos proporcionara resultados más acertados en la detección de mensajes de texto agresivos.

5.3 Detección de mensajes de texto agresivos

El conjunto de datos que se obtuvo de `Twitter` está compuesto por comentarios que contienen la palabra clave *school*. Esta palabra fue escogida, debido a que es un ambiente propenso al acoso e inclusive puede llevar a conversaciones que nos dirijan a este tipo de agresión. Los comentarios recolectados pertenecen al idioma inglés, así como los lexicones que utilizamos en nuestras propuestas de enfoques. Para trabajos futuros se considera trabajar con el idioma español.

De los comentarios que se consiguieron, solamente seleccionamos los comentarios que se encuentran dirigidos a una o más personas; debido a que el ciberacoso, como lo menciona la definición, es una agresión que es considerada hacia una víctima de manera directa. Con el repositorio filtrado por mensajes direccionados, se generaron dos conjuntos de datos. Uno de estos conjuntos está compuesto por comentarios que incluyan la palabra *f*ck* y el otro conjunto de datos contiene mensajes con la palabra *b*tch*.

La razón de escoger comentarios con malas palabras obedece a la intención de encontrar agresividad. Además se conoce que con estas dos malas palabras tienen un cierto grado de ambigüedad. El conocer los puntos de vista de los evaluadores sobre la amibigüedad de los comentarios, si estos los consideran agresivos o no y compararlos con los resultados de nuestra metodolgía nos percatamos si es considerado un mensaje agresivo o no a pesar de las palabras semillas que se escogieron. Un resumen de los conjuntos de datos obtenidos de `Twitter`, se pueden observar en la tabla 5.2. En la tabla 5.3 presenta ejemplos de cómo son los mensajes de texto que componen los conjuntos de datos.

Los conjuntos de datos fueron calificados por cuatro evaluadores,ver tabla 5.4. Su función consistía en calificar cada uno de los mensajes de texto, en una escala de cero a diez, tomando en consideración el cero como nada agresivo y el diez como muy agresivo.

La escala de agresividad que se está utilizando se considera adecuada debido a que nos apoya a deducir que tan agresivo puede llegar a ser un comentario. En la investigación que se realizó para hacer esta tesis, se pudo observar, en el estado del arte los distintos enfoques para detectar el ciberacoso entre ellos los valores binarios (Xu *et al.*, 2012) y la clasificación con polaridades en el análisis de sentimiento (Pang y Lee, 2004), por mencionar algunos.

Estas clasificaciones no ayudan a detectar si el mensaje cuenta con agresividad, solo apoya a decidir, según el criterio que consideraron, si existe ciberacoso o no. Como se ha mencionado anteriormente el ciberacoso consta de varias variables para determinar que es ciberacoso. Por ende, un valor binario y una polaridad que se va de un extremo a otro, no ayuda a decidir que tan agresivo es un mensaje dentro de una conversación. La intención de utilizar una escala de cero al diez, es para apoyarnos a decidir que tan agresivo es el mensaje considerando el contexto del mensaje.

Se escogió una escala del cero al diez, debido a que es una escala muy adecuada en cantidad de factores que la conforman (Diener *et al.*, 1985; Denscombe, 2014), no es tan extrema como del cero al cien y sobretodo una escala muy familiarizada para la evaluación con los evaluadores, para decidir el nivel de agresividad.

Tabla 5.3: Ejemplos de mensajes de texto.

Mensajes de texto
@F*ckCrystal i gotta go to school at 5 so if you wanna chill after that, bitch??
@ParishRory @Sh4niqua pff hah no way she is a fucking bully. I'm actually scared of her
@BitchnameLabre bitch I hate you yeah lol.
@BritishBitch I really do hate school haha.
@ASAPBACON171 lol bitch I'm good. I'm in school hoe.
@HotHeadG5T lol shuddup bitch but u still in school ?

Tabla 5.4: Ejemplo de una muestra del conjunto de datos y como la calificaron los evaluadores.

Comentarios	E1	E2	E3	E4
*@F*ckCrystal i gottta go to school at 5 so idk if you wanna chill after that?*	0	0	1	5
*@ParishRoy @5h4niqua pft hah no way she's a f*cking bully. Im actually scared of her*	4	7	5	8
*@coexisting f*ck school. High tuition prices are just a way to try to keep our generation in check by owing the government money forever*	0	4	5	5

Cada integrante del comité evaluador cuenta con las siguientes caracteristícas:

1. Ser mayores de edad, a causa de los mensajes de texto agresivos que se tengan que leer.
2. Deben dominar el idioma en el que se encuentre la base de datos con la que se va a trabajar. Para nuestro caso de investigación, el idioma inglés.

El proceso de como se llevó a cabo la calificación, a cargo de los evaluadores, se puede observar en la figura 5.2.

Este proceso se representa de la siguiente manera:

$c = [0..10]$ en donde, c es el mensaje de texto del conjunto de datos a evaluar y los evaluadores determinaron una calificación entre cero y diez.

Se clasificaron los mensajes de texto calificándolos por parte de cada evaluador. Con cada una de las calificaciones de los evaluadores se generó un promedio para cada mensaje. Con el promedio que se obtiene de cada mensaje se crea un nivel de agresividad manual. Cada evaluador generó sus calificaciones y se procedió a realizar una prueba estadística de análisis de varianza, para corroborar que los datos

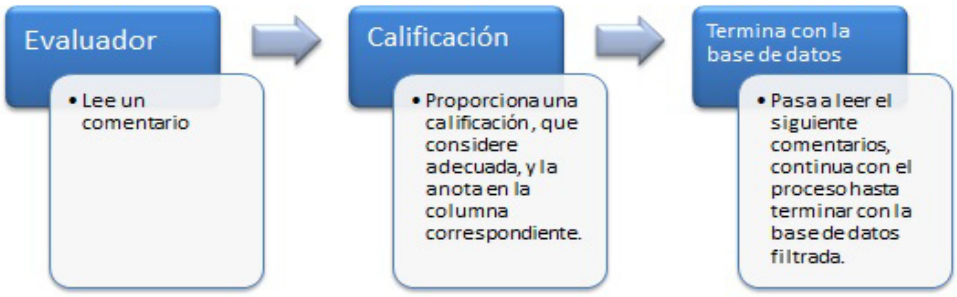

Figura 5.2: Proceso para elaborar la calificación por parte de los evaluadores

contaran con similitud entre ellos. Con estos resultados se generó un promedio al cual se le denominó << *promedio-humano*>>.

Para verificar que el criterio de los evaluadores fuera similar y que fuera útil para nuestros propósitos, utilizamos el análisis de varianza (ANOVA), ver tabla 5.5.

Algunos mensajes fueron descartados cuando se generó la prueba, dejando del conjunto de *f*ck* con 174 mensajes y el conjunto de *b*tch* con 69 mensajes. Los mensajes que se eliminaron fueron los que no contemplaban un juicio uniforme (por ejemplo, un evaluador colocó en un mensaje una ponderación leve, entre cero y cuatro y otro evaluador, en ese mismo mensaje, puso una ponderación grande, entre cinco y diez). Con esto se revela el grado de complejidad de la tarea, ya que incluso los humanos no están de acuerdo en un porcentaje de los mensajes.

Tabla 5.5: Resultados del análisis de varianza

ANOVA	conjunto de datos *f*ck*	conjunto de datos *b*tch*
Cantidad de mensajes (n)	174	69
Cantidad de evaluadores (a)	4	4
Subtotal de variable independiente [A]	4610.2	1946.1
Suma de [A] [T]	4275	1685.2
Valor Individual [Y]	14271	3868
Suma de cuadrados de grupos entre	334.9	260.9
Suma de cuadrados de grupos intra	9660.8	1921.9
Grados de libertad de grupos entre	3	3
Grados de libertad de grupos intra	692	272
Media cuadratica grupos entre	111.6	86.9
Media cuadratica grupos intra	13.9	7.1

Este análisis se realizó debido a que los insultos se pudieron percibir de maneras distintas, por diferentes personas (Goyal y Kalra, 2013); esto se refiere a que al leer el mismo mensaje de texto, cada evaluador consideró calificar el contenido del mismo con un valor de agresividad diferente y con ello dar realce a la investigación ya que no se omitió esta característica en el procedimiento.

Al tener los mensajes ya evaluados se procedió a preparar la información para que el programa generara el valor de agresividad de manera autómatica. Los mensajes de texto en los dos conjuntos de datos fueron previamente procesados. Debido a que se eliminaron los mensajes que tuvieran signos de puntuación, cada palabra se cambió a minúsculas y se utilizaron expresiones regulares. Para las expresiones regulares se corrigieron palabras mal escritas, por ejemplo: <<*biatch o biotchhhh*>>, las siglas se expandieron como <<*OMFG*>>, y se separaron las palabras de las malas palabras, por ejemplo un nombre de un usuario era *@muppybitch*, lo dividimos en *@muppy* y en *b*tch*. Además se tradujeron *emoticons* como :) , : (, y : @ por términos afectivos como *"happy"*, *"sad"*, *"angry"*.

Este procesamiento previo de la información se consideró para que los mensajes tuvieran una coherencia en la escritura. Además esta característica en el trabajo se consideró gracias a trabajos propuestos en el área del análisis de sentimiento como Bosse y Stam (2011), Bayzick *et al.* (2011) y Sanchez y Kumar (2011).

5.3.0.1 Lexicones utilizados en la experimentación

Dentro del estado del arte en el área del análisis de sentimiento, se encuentran listas de apoyo dentro del sentimiento de agresión o negativo. Para nuestros experimentos se utilizaron tres lexicones, ver tabla 5.6.

El primer lexicón fue del sitio `noswearing.com`, un sitio que almacena contribuciones por parte de la comunidad anglosajona de palabras ofensivas así como su significado. Dado que es la misma comunidad quienes aportan las palabras a este sitio, la lista tiene como beneficio que se encuentra compuesta por el *slang* y por palabras que van surgiendo con el tiempo (Sood *et al.*, 2012). Este lexicón es el que se menciona en el capítulo de metodología.

El segundo lexicón se obtuvo del estudio llamado: <<Affective Norms for English Words>> (ANEW). ANEW se formó por los participantes que evaluaron su reacción a un conjunto de 1034 palabras con respecto a tres estándares semánticos diferenciales de bueno-malo (valencia psicológica), activo-pasivo (motivación) y fuerte-débil (dominio). Se utiliza una escala del uno al nueve, donde uno es lo menor en la escala de cada uno de los estándares y nueve el mayor en cada uno de los estándares (Dodds y Danforth, 2010).

El tercer lexicón es SentiWordNet. SentiWordNet es una herramienta léxica para la minería de opinión. Esta herramienta utiliza la base de datos de WordNet, una base de datos que contiene las palabras en inglés. WordNet contiene nombres, verbos, adjetivos y adverbios agrupados en conjuntos de sinónimos cognitivos, llamados <<*synsets*>>. Cada *synset* expresa un concepto distinto. Los *synsets* están

vinculados entre sí por medio de las relaciones conceptuales semántico y léxico. SentiWordNet asigna a cada *synset* de WordNet tres clasificaciones con respecto a la confianza, la negatividad, la positividad y la objetividad (Ventura de Souza, 2011).

Cada *synset* se asocia a tres valores numéricos Pos(s), Neg(s) y Obj(s) que indican los términos positivos, negativos u objetivos (o neutros) están contenidas en cada *synset*; cada valor está dentro del intervalo [0.0, y 1.0] y la suma de los tres valores asociados es necesariamente 1.0. Esto significa que cada *synset* tiene un valor distinto de cero en al menos una de las categorías (Ventura de Souza, 2011).

Tabla 5.6: Lexicones de palabras que se utilizaron para la experimentación.

No.	Nombre de la lista	Cantidad de palabras en la lista
1	Noswearing.com	350
2	Affective Norms for English Words (ANEW)	1034
3	SentiWordNet	204,560

Cada uno de estos tres lexicones fue utilizado de manera individual, para cada conjunto de datos en el programa para generar el valor del nivel de agresividad a cada mensaje respectivamente. El proceso para generar el nivel de agresividad es similar. Se realizó un conteo de las palabras del mensaje de texto que se encuentran en el lexicón correspondiente. Lo que hace la diferencia entre cada lista son los valores que contienen las palabras de los lexicones, se pueden ver ejemplos de estos valores en la tabla 5.7. Es decir, para *SentiwordNet*, se sumaron los valores, de la clasificación negativa, de las palabras del mensaje encontradas en el lexicón.

La característica de *ANEW*, es que los valores que contienen las palabras de este lexicón no representan negatividad. Este valor refleja su valor de afectividad, es decir, qué tan amigable es considerada la palabra. Es por esto, que con *ANEW*, al igual que *SentiWordNet*, se sumaron los valores de las palabras del mensaje de texto que se encontraron en el lexicón. Se generó el valor para cada mensaje de texto como se mencionó en la metodología y al contar con este valor se generó su inverso. El inverso se obtuvo restando el valor mayor del estándar de *ANEW*, el cual es nueve,

menos el valor generado por nuestra metodología.

Por ejemplo, para calcular el valor del nivel de agresividad con ANEW, se tomaron los valores totales de las palabras encontradas en el lexicón y los promediamos; por ejemplo, dado un documento $d_i = \{w_1, w_2, w_3, w_4\}$, si w_2 y w_4 encontramos que su respectivo valor es 5.0 y 8.5, el valor promedio sería $\frac{(5.0+8.5)}{2} = 6.75$. Debido a que este valor refleja el grado de felicidad lo que hace que se incrementa los valores (lo contrario a nuestra escala, donde los valores más grandes son más negativos), como el rango de ANEW difiere al de nosotros, traducimos los promedios resultantes utilizando

$$\begin{aligned} sc_i &= \frac{(b-a)\left[(d-v_i)-c\right]}{d-c} + a \\ &= \frac{(10)\left[(9-v_i)-1\right]}{8} \end{aligned} \tag{5.1}$$

donde $a = 0$, $b = 10$, $c = 1$, $d = 9$, y v_i es el valor del promedio obtenido del documento d_i; notese que $[a, b]$ es nuestro rango de agresividad y $[c, d]$ es el rango de felicidad de ANEW. Para el ejemplo proporcionado anteriormente, el valor promedio de $v_i = 6.75$ se traduciría en valor de nivel de agresividad $sc_i = 1.56$.

La diferencia entre *Anew, SentiWordNet y Noswearing*, es que la lista de palabras de este último, no cuentan con valores determinados en sus palabras. Es por esto que para este lexicón se realiza el conteo como se menciona en la metodología.

5.3.0.2 Implementación de los sistemas difusos

Otro enfoque que se utilizó en la comparación de los métodos que sean factibles para la detección de agresividad en los mensajes de texto, fue la lógica difusa. La lógica difusa se considera equiparada a la computación con palabras. La computación con palabras es una metodología en la que las palabras se utilizan en lugar de los números para la informática y el razonamiento (Zadeh, 1996).

Tabla 5.7: Ejemplos de las palabras con sus respectivos valores que conforman los lexicones.

Lexicón	Palabra	Valor
ANEW	friendly	8.43
	fame	7.93
	frigid	3.5
	family	7.65
	frog	5.71
SentiWordNet	sensate	0.25
	insensate	0.5
	unfeeling	0.5
	animate	0.0

La herramienta de software que se utilizó para esta fase de la experimentación se llama <<qtfuzzylite>> `http://www.fuzzylite.com/`, la cual es una plataforma cruzada, es de código abierto basado en gráficos de interfaces de usuario. Su objetivo es permitir diseñar de manera visual los sistemas difusos de fuzzylite y poder interactuar con ellos en tiempo real (Rada Vilela, 2013).

Para este programa se requiere colocar datos de entrada, los datos de entrada que se utilizaron fueron: número de cantidad de palabras que componen el mensaje de texto y cantidad de palabras agresivas detectadas en el mensaje de texto. La variable de salida que nos arroja el programa es el nivel de agresividad. El sistema difuso que se utilizó fue el Mandami y el defuzzificador fue el centroide.

En la figura 5.3, se muestra una captura de pantalla en donde se puede observar, las entradas, las salidas y las reglas de inferencia.

Considerando nuestro patrón de los mensajes en el conjunto de datos y de cómo se encuentra la clasificación de estos conjuntos, para lograr lo que se pretende con lógica difusa se procede a realizar las reglas de inferencia, que forman nuestra base

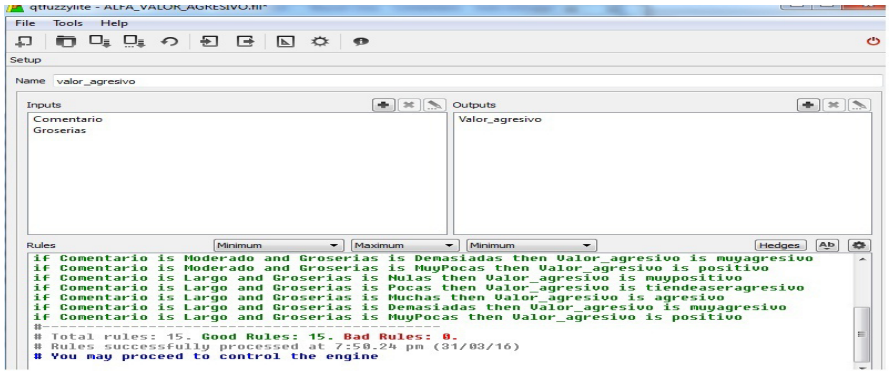

Figura 5.3: Captura de pantalla de qtfuzzylite

de conocimiento. Estas reglas se pueden observar en los apéndices A.3, A.4, A.5.

Las reglas de inferencia,como se puede ver en la tabla 5.8, se encuentran establecidas con los conjuntos difusos. Estos conjuntos se encuentran determinados para los conjuntos difusos de entrada los cuales son:

- La cantidad de palabras que forman el comentario.
- La cantidad de malas palabras que se encuentran en el comentario
- Para el conjunto difuso de salida, el cual genera el valor de agresividad.

La cantidad de palabras para determinar si un comentario era corto, moderado o largo, se basó en un promedio generado para cada posible agresor. Se contaban las palabras que contenían los comentarios de cada posible agresor y se determinó un promedio para de ahí determinar el criterio de los comentarios. Se realizó el mismo proceso para determinar si la cantidad de malas palabras o groserías son demasiadas, pocas o nulas.

En la tabla 5.9 se pueden observar los conjuntos difusos que se crearon para el desarrollo de este experimento con este enfoque de lógica difusa.

Tabla 5.8: Ejemplo de las reglas de inferencia que se utilizaron para el experimento

If Comentario is Moderado and Groserias is Demasiadas then Valor agresivo es muy agresivo
If Comentario is Corto and Groserias is Pocas then Valor agresivo es tiende a ser agresivo
If Comentario is Largo and Groserias is Nulas then Valor agresivo es muy positivo

Tabla 5.9: Conjuntos Difusos

Cantidad de palabras en el comentario	Cantidad de Malas Palabras	Valor Agresivo
Muy Corto	Nulas	Muy positivo
Corto	Muy pocas	Positivo
Moderado	Pocas	Tiende a ser agresivo
Largo	Muchas	Agresivo
Muy Largo	Demasiadas	Muy agresivo

5.3.1 Resultados

Para la comparación, con cada uno de los procedimientos, se utilizó el error medio cuadrático (MSE), el cual es calculado como $(x - y)^2$, donde x es el valor del promedio humano y y es el valor obtenido por cada uno de los enfoques.

Para tener una vista clara de los resultados, se introdujo un *baseline*. El *baseline* viene siendo una referencia, como un punto de partida básico que nos apoya en determinar si nuestro proceso es adecuado. Esta referencia consiste en valores generados aleatoriamente. Estos valores fueron generados treinta veces y fueron promediados.

Nuestros resultados se muestran en las figuras 5.4, 5.5 y 5.6 ; en la tabla 5.10, se encuentran los datos que representan los resultados de los errores cuadráticos medios *(MSE)*.

El enfoque que tuvo mejor resultado, el que tiene un *MSE* bajo, fue la lógica difusa, seguido por el *lexicón* de *No Swearing*, despúes por el *lexicón* de *SentiWordNet* y finalmente por el *lexicón* de *ANEW*.

Fue interesante encontrarse que el lexicón *ANEW*, fue un enfoque que salió con un error más alto que el *baseline*; creemos que puede deberse a la presencia de *slang* y de texto informal, así como alguna ambigüedad.

Si comparamos los lexicones con la lógica difusa, se encuentra una diferencia importante. Además, si comparamos los resultados obtenidos por cada conjunto de datos, podemos observar que el conjunto de datos con f*ck en general obtiene un *MSE* más bajo que el conjunto de datos con b*tch; esto se puede deber a el tamaño del conjunto de datos.

Dentro de los enfoques de los lexicónes, los mejores resultados se obtuvieron por parte de *NoSwearing*; si bien esto podría ser, en parte, a los conjuntos de datos (elegido por la búsqueda de malas palabras), creemos que la fuerza del enfoque más bien radica en la estrecha relación que existe entre la agresividad y lenguaje profano. En ese sentido, la presencia y el número de malas palabras en el texto pueden actuar como una característica clave para la detección de agresividad en el texto y con ello alcanzar la detección de casos de ciberacoso.

Tabla 5.10: Resultados MSE

Enfoque	conjunto f*ck	conjunto b*tch	Promedio
NS	5.2	7.2	6.23
ANEW	16.1	33.9	24.95
SentiWordNet	11.2	8.4	9.8
Lógica Difusa	4.8	6.1	5.5
Baseline	15.6	20.1	17.9

Analizando los resultados obtenidos, también es importante tener en cuenta que los casos más difíciles de detectar para todos los enfoques eran los que tienen un alto grado de agresividad; creemos que esto se debe a varias razones. Por otra parte, podría haber aspectos en los comentarios que deben considerarse, tales como emociones subyacentes, intenciones y el contexto.

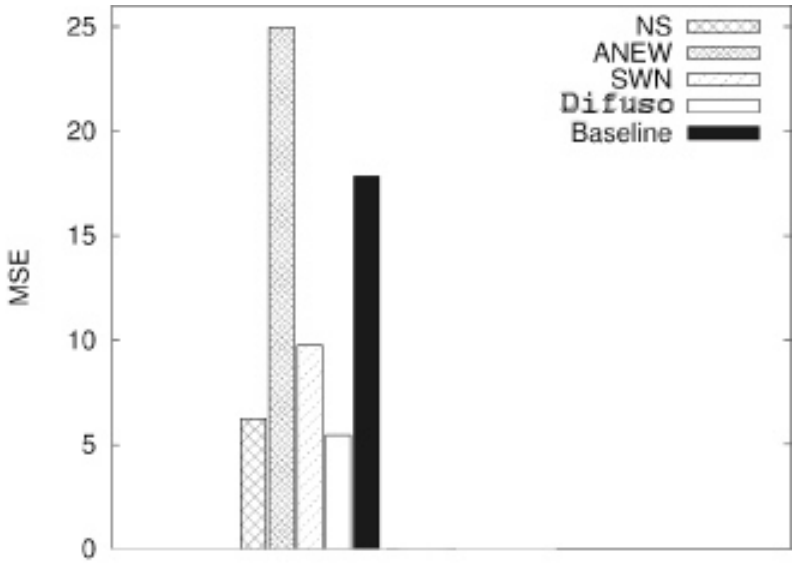

Figura 5.4: Promedios del error cuadrático medio (MSE). NS= `noswearing.com` lexicón, SWN= SentiWordNet.

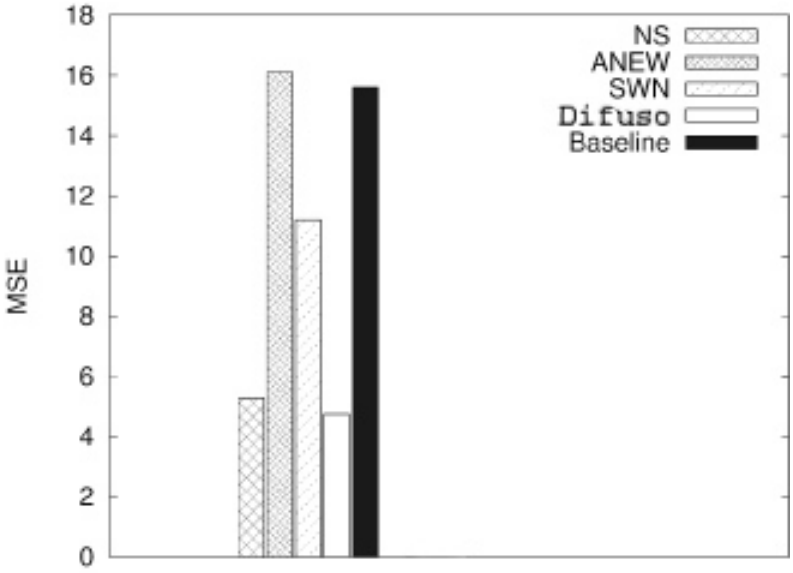

Figura 5.5: Error cuadrático medio por conjuntos de datos con f*ck.

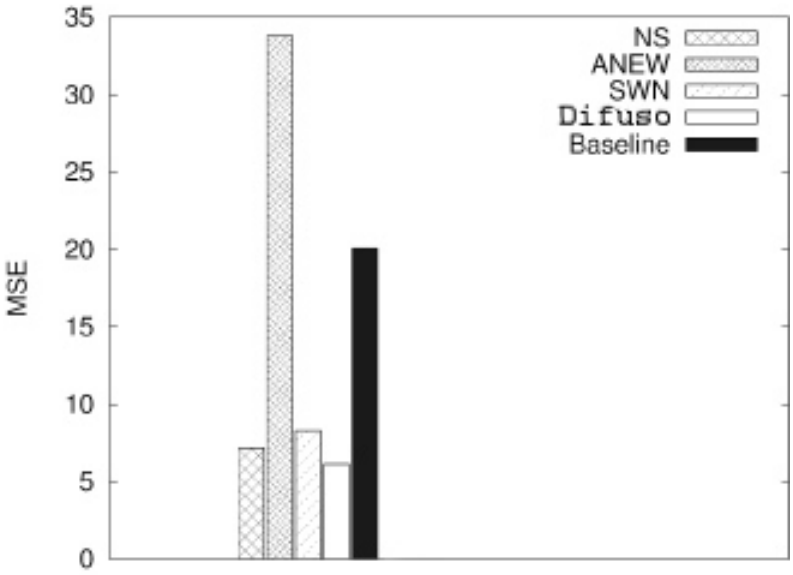

Figura 5.6: Error cuadrático medio por conjuntos de datos con b*tch.

5.3.2 Discusión general de los resultados

En el presente trabajo, hemos abordado la detección del texto agresivo que consiste en mapear un documento con un valor de agresividad; como una mejora para nuestro trabajo, en el estado del arte, los métodos que se encuentran tienden a emitir este tema como un problema de clasificación binaria. Hemos definido una simple escala que va de cero a diez (donde diez es el más agresivo) y supone, además, que la detección del texto agresivo es una subtarea dentro del área del análisis de sentimiento que está estrechamente relacionado con la detección de polaridad en documentos.

Tomando en cuenta lo anterior, propusimos y exploramos los siguientes métodos: lexicones y lógica difusa. Los métodos se pusieron a prueba sobre conjuntos de datos obtenidos de `Twitter`. Nuestros resultados muestran que la lógica difusa y el lexicón de *NoSwearing* son candidatos sólidos para evaluar los conjuntos de datos. Además que el uso de lenguaje profano (malas palabras) también parece ser una característica clave para la tarea. Es por esta razón que se considera factible el utilizar

el lexicón de *NoSwearing* en nuestros experimentos. Los resultados obtenidos con este lexicón, son buenos.

El uso del lexicón *Noswearing* ha sido seleccionado sobre la lógica difusa, debido a que el lexicón genera los datos de forma automática, sobre el proceso que requiere más manejo manual en la lógica difusa.

5.4 Detección de casos de ciberacoso

En esta sección se explican los experimentos realizados para lograr detectar los casos de ciberacoso. A diferencia de la detección de los mensajes agresivos, en esta tarea se cuenta con la ventaja de que ya contamos con un proceso adecuado, experimentado y con resultados de lo que es un mensaje agresivo. Teniendo la información de como es el proceso para la detección de mensajes agresivos, se prosigue a generar los experimentos para lograr la detección de los casos de ciberacoso, en base a la definición de ciberacoso mencionada en la metodología.

5.4.1 Configuración

Para estos experimentos se utilizó el conjunto de datos con el que se trabajó para la detección de mensajes agresivos, pero además se integraron más datos al conjunto con la finalidad de tener un número de casos suficientes para el desarrollo de experimentos.

Se utilizó el proceso de filtrado que se mencionó para la detección de agresividad en los mensajes, como se explica en la metodolgía 4.1. Además se utilizaron diferentes palabras agresivas como clave para generar más datos, dichas palabras se muestran en la tabla 5.11. Con estas palabras clave más los datos que ya contábamos se generó un conjunto de datos conformado por $13,313$ mensajes direccionados en el idioma inglés. El nuevo conjunto de datos generado por la extensión de palabras claves lo

definiremos como B_1.

Tabla 5.11: Palabras agresivas que se emplearon para generar más datos para nuestra investigación

*c*nt*	*wh*re*
punk as b*tch*	*b*tch as* nigg*r*
punk as nigg*er*	*f*ggot*
*f*ck*ng f*ggot*	*f*ck*ng sl*t,*
*f*ck*ng c*nt*	*motherf*cker*

Al tener identificados los agresores, se procedió a buscar dentro de la red social los mensajes que ha realizado agresor por agresor en un período determinado; en nuestro caso fue de seis meses. Además de los mensajes enviados, se obtuvo la fecha en la que se envió el mensaje, así como el receptor. Con esto se genera una base de datos a la cual denominaremos B_2.

Este proceso se realizó con diez agresores y con sus víctimas correspondientes para cada caso los cuales iban conformando B_2.

Al término de procesar la información y de detectar los casos de ciberacoso, se prosigue con la comparación de los resultados obtenidos por nuestro procesamiento de datos y con la evalaución manual por parte de evaluadores. Las características de los evaluadores son las mismas que se manejaron para la etapa anterior de detección de mensajes agresivos.

La evaluación manual se realizó de una manera muy similar a como se realizó con la detección de mensajes agresivos:

- Se les proporcionó a nueve evaluadores una base de datos, a la que nombramos: *encuesta*.
- Se les solicitó a cada uno de los evaluadores que leyeran los mensajes de la

encuesta y colocaran, si a su percepción, existe un caso de ciberacoso o no.

- Esta encuesta esta conformada por veintiseis casos. Estos casos están compuestos por mensajes de conversaciones, provenientes de diez agresores con sus víctimas. Cada caso es diferente, es decir, no se repiten el agresor con su víctima.

- Cabe mencionar que los casos se obtuvieron de B_2, para asegurarnos de colocar casos detectados como ciberacoso por nuestra metodología. Así como también, se agregaron intencionalmente casos que no cuentan con agresividad o que se identificaron claramente como casos que no forman parte de un ciberacoso.

- Esto se realizó porque la finalidad de este experimento es la de comprobar que los casos detectados como ciberacoso por nuestra metodología son también considerados casos de ciberacoso para una persona común.

La forma de como se iban etiquetando, de manera global, cada uno de los comentarios por parte de los evaluadores, es en base a la opinión de los resultados de la mayoría. Es decir, si en un comentario ocho evaluadores mencionan que sí es un caso de ciberacoso y dos califican lo contrario, el comentario es considerado ciberacoso.

5.4.2 Resultados

Los resultados de la comparación manual junto con los obtenidos por nuestra metodología se pueden observar en la tabla 5.14. De estos resultados se obtuvo una medida de F de 95.99. Para seguir corroborando resultados se realizó una matriz de confusión, los resultados se encuentran en la tabla 5.13 y en la figura **??**. Siendo la figura **??** una representación gráfica de la tabla 5.14.

Tabla 5.12: Resultados de conjuntos de datos obtenidos de `Twitter` para detectar casos de ciberacoso

Conjunto de datos	Cantidad de mensajes en el conjunto
B_1	13,313

Tabla 5.13: Resultados de matriz de confusión para detección de casos de ciberacoso

	Casos Positivos	Casos Negativos	Total
Casos Positivos	**18**	2	20
Casos Negativos	0	**6**	6
Total	18	8	**26**
Número de clasificaciones correctas : 24 (92 %)			

Tabla 5.14: Resultados de la metodología propuesta en esta investigación para detección de casos de ciberacoso

Casos Reales (R)	Casos Detectados (D)	Casos Reales Detectados (RD)	% Precisión (RD/D)	% Exhaustividad (RD/R)	F $_{(2*[(P*E)/(P+E)]}$
26	24	24	100	92.30	95.99

5.4.3 Discusión general de los resultados

Como se puede observar en la tabla 5.13 los casos con resultados detectados como ciberacoso por nuestra metodología son veinte y los casos que no son detectados como ciberacoso por nuestra metodología son seis. Los resultados de casos detectados como ciberacoso por nuestra metodología y por los evaluadores fueron dieciocho. Los

resultados detectados como ciberacoso por nuestra metodología como positivos pero los evaluadores no los detectaron como casos de ciberacoso fueron dos.

Analizando los dos casos, que se pueden observar en la tabla 5.15, se puede deducir las razones por las que el resultado por parte de los evaluadores haya salido diferente a el de la metodología. La primera razón es que de los nueve evaluadores, cinco evaluadores (mayoría), calificaron como que no son casos de ciberacoso. Tomando estos dos casos como negativos cuando nuestro proceso los calificó como casos de ciberacoso, es decir, los consideró positivos.

La metodología consideró positivos estos casos debido a las palabras ofensivas que se encuentran en los mensajes. Además que, cabe mencionar, que para la detección del caso de ciberacoso se considera un promedio de la agresividad de todos los mensajes, provenientes del agresor con la víctima. En la encuesta solo se considera una parte de los mensajes para que la encuesta no se volviera pesada para los evaluadores. Esto se puede observar en la sección de apéndice . Así como también evitar el posible cansancio en los mismos y que esto ocasionara un sesgo en las respuestas.

Tabla 5.15: Ejemplos de casos utilizados para evaluación manual

Agresor	*Rathgrith027*
Víctima	*andymanhands*
Mensajes	*@andymanhands Simple. Every fucking fuckfest made in RPG Maker VX. @andymanhands even know the latter was even possible, but yet, here we are at your house at the intersection of clusterfuck road and jackass lane. @andymanhands Nah, I'm kekking all the way to the bank. Morron*
Caso de Ciberacoso	Positivo para la metodología
Agresor	*ruv1nay*
Víctima	*Byoncaa*
Mensajes	*@Byoncaa yeah my nigga. @Byoncaa Jesus Christ. @Byoncaa lmao man. @Byoncaa says the freshman. @Byoncaa because freshman's are dumb*
Caso de Ciberacoso	Positivo para la metodología

5.5 Resumen

El capítulo de experimentos y resultados presentó las actividades realizadas para esta investigación. Además presentó las soluciones obtenidas en base a los experimentos desarrollados. Este capítulo se dividió en dos secciones: detección de mensajes de texto agresivos y detección de casos de ciberacoso. Cada sección contiene tres subsecciones: configuración, resultados y discusión general de los resultados.

En la sección correspondiente a la detección de mensajes de texto agresivo, se

desarrollaron los experimentos que se utilizaron para generar el nivel de agresividad, para cada mensaje de texto. Los mensajes de texto se obtuvieron de la red social `Twitter` y con estos mensajes se formaron dos conjuntos de datos. Un conjunto de datos formado por mensajes que contenian la palabra *f*ck* y otro conjunto con mensajes que contenían la palabra *b*tch* .

Para la detección de mensajes agresivos se realizó una comparación entre la evaluación que se obtuvo para cada mensaje de los conjuntos de datos por parte de tres lexicones: *Noswearing*, *Anew* y *SentiWordnet*. Igualmente se empleó lógica difusa para que se uniera a la competencia con los lexicones. Para comprobar la eficacia de nuestros métodos al igual se generó una base como referencia, compuesto por números generados aleatoriamente. Los enfoques que tuvieron mejores resultados fueron la lógica difusa y el lexicón noswearing.

Para comprobar que los resultados obtenidos por nuestra metodología se asemejaban a lo que un ser humano considera un mensaje agresivo, se realizó una evaluación manual por parte de cuatro evaluadores. El producto de esta investigación mostró los casos de como la comparación entre las listas de palabras, que se ocupan en el estado del arte para detectar sentimiento y la calificación de lo que pensaron los evaluadores de los comentarios utilizados en las pruebas, dan resultados semejantes.

Con el proceso para la detección de mensajes agresivos, se procedió a desarrollar la detección de casos de ciberacoso. Para este proceso se utilizaron más palabras agresivas como por ejemplo: *c*nt*, *wh*re*,*f**gg*t*, entre otras, ver tabla 5.11. Se generó un conjunto de datos con estas palabras y se procedió a detectar quienes son los emisores de los mensajes que cuentaban con un nivel de agresividad mayor de cinco. Estos emisores se consideraron los <<agresores>>.

Teniendo detectados a los agresores, se realizó una búsqueda de las conversaciones de cada uno de estos agresores, en un período de seís meses. Se filtraron los mensajes agresivos de estas conversaciones y se detectaron las <<víctimas >>.

Al contar con los agresores y víctimas detectados, se generó el nivel de agresi-

vidad para cada mensaje emitido por el agresor hacia cada una de sus víctimas. Se realizó una suma para cada una de las conversaciones con sus respectivas víctimas. Contando con estas sumas se generó un promedio de agresividad. La conversación que contara con una agresividad mayor al promedio es considerado un caso de ciberacoso.

Para comprobar que los resultados obtenidos por nuestra metodología se asemejan a lo que un ser humano consideraba un mensaje agresivo, se realizó una evaluación manual por parte de nueve evaluadores, tal como se utilizó en para la detección de mensajes agresivos. El producto de esta investigación muestra los casos de como la comparación entre nuestro proceso y la calificación de lo que piensan los evaluadores de los comentarios utilizados en las pruebas, dieron resultados semejantes.

Capítulo 6

Conclusiones y trabajo futuro

Nunca consideres el estudio como una obligación, sino como la oportunidad para penetrar en el bello y maravilloso mundo del saber.

Albert Einstein

Este trabajo de investigación presenta una manera de poder detectar ciberacoso de manera automática en redes sociales por medio de herramientas que se encuentran dentro de las tecnologías de información como lo es el análisis de sentimiento a través de la clasificación de texto.

En el Capítulo 4 se presenta una metodología la cual nos apoya a encontrar casos de ciberacoso en una red social; la metodología está compuesta por tres etapas, las cuales son definidas por el concepto de ciberacoso utilizado para esta investigación.

La metodología de este estudio consiste en encontrar mensajes considerados como agresivos, colocar un valor de agresividad a cada uno de estos mensajes, analizar el usuario que envía estos mensajes considerados agresivos para proporcionarle a el usuario un valor de agresividad y teniendo el nivel de agresividad con el que se

comunica el usuario se revisa sus conversaciones para poder encontrar conversaciones realizadas de manera frecuente y consideradas agresivas en un período de tiempo. Si esto es positivo se considera un caso de ciberacoso.

Este manera de poder detectar el ciberacoso genera datos e información que con ello ayudan al estudio dentro de esta área de investigación tanto social como de manera tecnológica.

En el Capítulo 5 se detallan los experimentos que se llevaron a cabo para lograr la detección de los casos de ciberacoso según la definición manejada en este trabajo en una red social. A su vez, se muestran los resultados obtenidos en cada una de las etapas definidas en la metodología propuesta para la resolución del problema planteado.

Con respecto a los experimentos realizados y a los resultados obtenidos se llega a la conclusión que el resultado de esta investigación mostró los casos de cómo la comparación entre los experimentos realizados con las listas de palabras, que se ocupan en el estado del arte para detectar sentimiento *(lexicones)* y la evalaución ponderada según del pensamiento de los evaluadores hacia los comentarios que se les presentaron y que fueron utilizados en las pruebas, dan resultados semejantes.

Con resultados semejantes nos referimos a que lo que un evaluador considera que el ejemplo que se le presentó, si es un caso de ciberacoso, nuestra solución propuesta concuerda a que si considera que ese ejemplo es un caso de ciberacoso.

6.1 Preguntas de investigación

En esta investigación, se generaron las siguientes preguntas:

- ¿Es posible detectar el ciberacoso en las redes sociales a través del análisis de sentimiento?

- ¿ Es posible generar un valor, un apoyo numérico y que con esto sea sencillo tomar una decisión que ayude a detectar agresividad en textos de una manera autómatica?
- ¿Se puede detectar de manera automática el ciberacoso a través de los textos?

La respuesta a estas preguntas es positiva, ya que como se observa en la investigación si es posible detectar el ciberacoso en una red social a través de herramientas que se utilizan dentro del área del análisis de sentimiento. Y que a su vez se genera un valor de agresividad el cual ayuda a poder detectar de una manera visual el que tanto es agresivo un comentario apoyandose en la escala de agresividad generada gracias al análisis de sentimiento. Y con los resultados que se obtuvieron en la investigación si se puede detectar de manera automática el ciberacoso a través de texto ya que en esto se fundamentó esta tesis, en mensajes de texto plasmados en una red social.

6.2 Contribuciones de la tesis

La contribución de esta tesis se basa en:

- Definir un nivel de agresividad el cual identifica de una manera adecuada la agresividad de un mensaje obtenido de una red social, considerando las características de la definición de ciberacoso y con ello apoya en la toma de decisión si el mensaje es agresivo o no, ya que permite decir qué tanto es lo agresivo del mensaje.
- Identificar a los participantes de las conversaciones que contengan los mensajes agresivos obtenidos de manera dinámica con el nivel de agresividad generado.
- Determinar la frecuencia con la que se envían los mensajes agresivos de los posibles agresores hacia las posibles víctimas, considerando que ya se cuenta con esta información.

Por lo tanto teniendo estos factores: la detección de mensajes agresivos, la detección de los participantes de las conversaciones (el agresor y receptor) y la frecuencia de este envío de mensajes por parte de los participantes, se obtiene la detección del ciberacoso; como se ha mencionado anteriormente la definición del ciberacoso es un acto agresivo (envío de mensajes agresivos), hacia una víctima (participantes), utilizando vías electrónicas y medios de comunicación electrónica (redes sociales) de manera consecutiva (frecuencia).

6.3 Trabajo futuro

Con el rumbo que esta investigación fue generando, surgieron otras cuestiones que no se encontraban consideradas en el estudio en sus inicios.

Una de estas cuestiones es a través de cómo se puede detectar el ciberacoso por medio del uso de imágenes y videos, una manera de poderlo resolver es utilizando técnicas de visión computacional y como se pueden aunar estas técnicas con las herramientas que utiliza el análisis de sentimiento. Además se deja como trabajo a futuro el utilizar métodos supervisados buscando métodos de muestreo y de ensamble que nos apoyen a identificar mensajes agresivos, todo esto tal como se mencionó en el Cápitulo 4. Como trabajo a futuro también se plantea el trabajar con el idioma español como se considera en el Cápitulo 5.

Otra pregunta que surgió es como utilizar de una manera más inclusiva la inteligencia artificial, en este lazo de estudio. En este caso, tratando de que sea por trabajos no supervisados como en el caso de la lógica difusa o en su defecto de poderlo manejar de una manera más automatizada.

En cuestiones académicas, es un área que puede explotarse en gran manera, ya que es un tema que a los estudiantes les llama mucho la atención, por el aporte social y tecnológico que se logra generar.

Dentro de este trabajo de investigación para tesis que se estuvo realizando por casi 3 años y medio, se ayudó a investigaciones que requirieron información sobre el tema del ciberacoso, sobre todo las bases de datos ya que si fue algo complicado conseguir este tipo de información.

6.4 Aplicaciones

Esta investigación ha apoyado en el inicio de proyectos el cual me ha permitido trabajar con los estudiantes y que ellos también se vean favorecidos con las aportaciones que se están generando. Uno de los proyectos en los que se va a empezar a trabajar es en realizar una aplicación para poder aplicar en redes sociales un tipo de alerta para que al momento de enviar un mensaje, este mensaje sea evaluado y en caso de que cuente con un grado de agresividad se le notifique al emisor en forma de pregunta si en verdad se encuentra seguro de enviar el mensaje agresivo, con este tipo de aplicación se puede encontrar de una manera más sencilla los mensajes agresivos ya que se le puede asignar un bit para identificar estos mensajes enviados y al igual que se puede considerar una intencón de agresividad por parte del emisor, debido a que éste decide enviar el mensaje a pesar de conocer que en verdad está siendo agresivo.

Apéndice A

Resultados de encuestas con evaluadores

Tabla A.1: Resultados por parte de los evaluadores para la detección de casos de ciberacoso.

	Resultados de cada uno de los evaluadores								
No. caso	1	2	3	4	5	6	7	8	9
1	S	S	S	S	S	S	S	S	S
2	S	S	S	S	S	S	S	S	S
3	N	N	N	N	N	N	N	N	N
4	N	N	N	N	N	N	N	N	N
5	S	S	S	S	S	S	S	S	S
6	S	S	N	S	S	N	S	S	N
7	N	N	N	N	N	N	N	N	S
8	S	S	S	S	S	S	S	S	S
9	S	S	S	S	S	S	N	S	N
10	S	S	S	S	S	S	S	S	S

Tabla A.2: Resultados por parte de los evaluadores para la detección de casos de ciberacoso

	Resultados de cada uno de los evaluadores								
No. caso	1	2	3	4	5	6	7	8	9
11	S	S	S	S	S	S	S	S	S
12	N	N	N	N	N	N	N	N	N
13	S	S	N	S	S	S	S	N	S
14	S	S	S	S	S	S	S	S	S
15	N	N	N	N	N	N	N	S	N
16	S	S	S	S	S	S	S	S	S
17	S	S	S	S	S	S	S	S	S
18	S	N	N	S	S	N	N	N	S
19	S	S	S	S	S	S	N	S	N
20	S	S	N	S	S	S	N	N	S
21	N	S	S	S	S	S	S	S	N
22	S	S	S	S	S	S	S	S	S
23	S	S	N	S	S	S	S	S	N
24	S	S	S	S	S	S	N	S	S
25	S	N	N	N	S	N	N	S	S
26	N	N	N	N	N	N	N	N	N

Tabla A.7: Ejemplo de encuesta que se le colocó a los evaluadores para detectar casos de ciberacoso.

No. caso	Agresor	Víctima	Tweets enviados por el agresor	Existe ciberacoso?S=Sí, N=No
1	BCAKESX	@V0RBY	@V0RBY technically those assists should be kills soo @V0RBY do it pussy @V0RBY I don't see what you're trying to do here motherfucker @V0RBY no xbox atm you retarded prick	
2	BCAKESX	@zumuhNE	@zumuhNE later motherfucker @zumuhNE i want to tit fuck you so hard @zumuhNE its just business baby, you know you're my one and only prick @zumuhNE aw shit beast fam fucker @zumuhNE hey zumuh whats up retarded prick @zumuhNE take it out my interactions boys gays	
3	GARY	@xobelmiic	@xobelmiic dont like white girls @xobelmiic am i purty @xobelmiic your eyes are crying in tears of happiness of how sexy king patel is @xobelmiic LOL	

Tabla A.8: Ejemplo de encuesta que se le colocó a los evaluadores para detectar casos de ciberacoso.

No. caso	Agresor	Víctima	Tweets enviados por el agresor	Existe ciberacoso?S=Sí, N=No
4	FUCKFORWARDS	@Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC	@Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC JADED IDIOTS! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC GAY HERBS! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC DEPRESSED DUMBASS HONKEYS! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC SUICIDAL IMBECILES! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC GAYLORD PRICKS! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC WEAK ASS HERB HOMOS! @Galileo908 @OfficialCarney @findjono @D00Z @ELLENandMrC GROW UP YOU IDIOTIC FAGBOY HONKEYS!	

Tabla A.3: Reglas de inferencia

No.	Comentario	Malas Palabras	Valor Agresivo
1	Si Comentario es muy corto	y Malas palabras son Nulas	entonces es positivo
2	Si Comentario es muy corto	y Malas palabras son muy pocas	entonces tiende a ser agresivo
3	Si Comentario es muy corto	y Malas palabras son Pocas	entonces es tiende a ser agresivo
4	Si Comentario es muy corto	y Malas palabras son Muchas	entonces es agresivo
5	Si Comentario es muy corto	y Malas palabras son demasiadas	entonces es muy agresivo
6	Si Comentario es corto	y Malas palabras son Nulas	entonces es muy positivo
7	Si Comentario es corto	y Malas palabras son muy pocas	entonces tiende a ser agresivo
8	Si Comentario es corto	y Malas palabras son Pocas	entonces es agresivo
9	Si Comentario es corto	y Malas palabras son Muchas	entonces es muy agresivo
10	Si Comentario es corto	y Malas palabras son demasiadas	entonces es muy agresivo

Tabla A.4: Continuación de las Reglas de inferencia

No.	Comentario	Malas Palabras	Valor Agresivo
11	Si Comentario es moderado	y Malas palabras son Nulas	entonces es muy positivo
12	Si Comentario es moderado	y Malas palabras son muy pocas	entonces es positivo
13	Si Comentario es moderado	y Malas palabras son pocas	entonces es tiende a ser agresivo
14	Si Comentario es moderado	y Malas palabras son muchas	entonces es agresivo
15	Si Comentario es moderado	y Malas palabras son demasiadas	entonces es muy agresivo
16	Si Comentario es largo	y Malas palabras son Nulas	entonces es muy positivo
17	Si Comentario es largo	y Malas palabras son muy pocas	entonces es positivo
18	Si Comentario es largo	y Malas palabras son pocas	entonces es tiende a ser agresivo
19	Si Comentario es largo	y Malas palabras son muchas	entonces es agresivo
20	Si Comentario es largo	y Malas palabras son demasiadas	entonces es agresivo

Tabla A.5: Continuación de las Reglas de inferencia

No.	Comentario	Malas Palabras	Valor Agresivo
21	Si Comentario es muy largo	y Malas palabras son Nulas	entonces es muy positivo
22	Si Comentario es muy largo	y Malas palabras son muy pocas	entonces es positivo
23	Si Comentario es muy largo	y Malas palabras son pocas	entonces es tiende a ser agresivo
24	Si Comentario es muy largo	y Malas palabras son muchas	entonces es agresivo
25	Si Comentario es muy largo	y Malas palabras son demasiadas	entonces es muy agresivo

Tabla A.6: Ejemplos de conversaciones de posible casos de ciberacoso.

Agresor	Víctima	Mensajes	Nivel de agresividad	Fecha
Ruv1nay	@HiiiGustavo	fat ass	5	Mayo 14
Ruv1nay	@HiiiGustavo	fucking faggot shit bruh	3.75	Mayo 14
Ruv1nay	@HiiiGustavo	sup wetback	5	Mayo 1
Ruv1nay	@HiiiGustavo	true my nigga	3.75	Mayo 15
Ruv1nay	@HiiiGustavo	DUMBASS WETBACK IM CALLING THE FBI	4	Abril 16
Ruv1nay	@HARLEYTHEGUY	pussy shit	5	Abril 24
Ruv1nay	@HARLEYTHEGUY	she beat your ass	2.5	Abril 30
Ruv1nay	@HARLEYTHEGUY	nigga got a black eye after	1.6	Abril 20
Ruv1nay	@salgadokelvin96	wetback	7.5	May 24
Ruv1nay	@salgadokelvin96	you just a faggot bruh get over it dick lover	8	Mayo 15
Ruv1nay	@salgadokelvin96	dumbass	7.5	Mayo 2
Ruv1nay	@salgadokelvin96	wetback	7.5	Abril 7
Ruv1nay	@salgadokelvin96	WITH BIG DICKS	5	Abril 20

Bibliografía

Abel, F., N. Henze, E. Herder y D. Krause (2010), «Linkage, aggregation, alignment and enrichment of public user profiles with Mypes», en *Proceedings of the 6th International Conference on Semantic Systems*, ACM, pág. 11.

Al-garadi, M. A., K. D. Varathan y S. D. Ravana (2016), «Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network», *Computers in Human Behavior*, **63**, págs. 433–443.

AOL (2015), «AOL Safe Chat», URL `http://www.aol.com/aim`.

Argamon, S., M. Koppel, J. Fine y A. R. Shimoni (2003), «Gender, genre, and writing style in formal written texts», *Text-the Hague then Amsterdam then Berlin*, **23**, págs. 321–346.

Ayala, D. V., E. Castillo, D. Pinto, I. Olmos y S. León (2012), «Information Retrieval and Classification based Approaches for the Sexual Predator Identification», en *CLEF (Online Working Notes/Labs/Workshop)*.

Baeza-Yates, R., B. Ribeiro-Neto *et al.* (1999), *Modern information retrieval*, tomo 463, ACM press New York.

Bayzick, K. April y E. Lynne (2011), «Detecting the Presence of Cyberbullying Using Computer Software», .

Berry, M. W. y M. Castellanos (2004), «Survey of text mining», *Computing Reviews*, **45**(9), pág. 548.

BHUTANI, A., D. K. MISRA y S. TOSHNIWAL (2012), «Detecting Socially Offensive Comments», *Indian Institute of Technology*.

BOOLE, G. (1854), *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*, Dover Publications.

BOSSE, T. y S. STAM (2011), «A normative agent system to prevent cyberbullying», en *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, tomo 2, IEEE, págs. 425–430.

BRASSLER, A. y O. HOMBURG (1996), «Integration of the fuzzy sets theory in the firms planning process», en *Proceedings of International Conference on Intelligent Technologies in Human-Related Sciences, León, Spain*, tomo 1, págs. 395–402.

BRAVO, C. B. y F. A. RASCO (2013), «Interacciones de los jóvenes andaluces en las redes sociales», *Comunicar*, **20**(40), págs. 25—30.

BREAZEAL, C. (2003), «Emotion and sociable humanoid robots», *International Journal of Human-Computer Studies*, **59**(1), págs. 119–155.

BUSSINES GUIDE INC, E. I. (2015), «The eBussines guide», URL `http://www.ebizmba.com`.

CAMPBELL, M. A. (2005), «Cyber bullying: An old problem in a new guise?», *Australian journal of Guidance and Counselling*, **15**(1), págs. 68–76.

CAÑELLAS, A. J. C. y L. B. BRAGE (2006), «Lógica difusa: una nueva epistemología para las Ciencias de la Educación», *Revista de educación,Ministerio de Educación. Centro de Publicaciones*, (340), págs. 995–1008.

CARMAGNOLA, F. y F. CENA (2009), «User identification for cross-system personalisation», *Information Sciences*, **179**(1), págs. 16–32.

CASAS, J. A., R. DEL REY y R. ORTEGA-RUIZ (2013), «Bullying and cyberbullying: Convergent and divergent predictor variables», *Computers in Human Behavior*, **29**(3), págs. 580–587.

CASTELLS, M. (2016), «Comunidades virtuales o sociedad red», .

CHEN, Y., L. ZHANG, A. MICHELONY y Y. ZHANG (2012), «4Is of social bully filtering: identity, inference, influence, and intervention», en *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, págs. 2677–2679.

CHISHOLM, J. F. (2006), «Cyberspace violence against girls and adolescent females», *Annals of the New York Academy of Sciences*, **1087**(1), págs. 74–89.

CYBERSAFETY (2015), «CyberSafety», URL `http://www.cyber-safety.com/`.

DADVAR, M. y F. DE JONG (2012), «Cyberbullying detection: a step toward a safer Internet yard», en *Proceedings of the 21st international conference companion on World Wide Web*, ACM, págs. 121–126.

DADVAR, M., F. DE JONG, R. ORDELMAN y R. TRIESCHNIGG (2012), «Improved cyberbullying detection using gender information», *Human Media Interaction Group*.

DEL REY, R., J. A. CASAS y R. ORTEGA (2015), «The impacts of the CONRED Program on different cyberbulling roles», *Aggressive behavior*.

DENSCOMBE, M. (2014), *The good research guide: for small-scale social research projects*, McGraw-Hill Education (UK).

DIENER, E., R. J. LARSEN, S. LEVINE y R. A. EMMONS (1985), «Intensity and frequency: dimensions underlying positive and negative affect.», *Journal of personality and social psychology*, **48**(5).

DINAKAR, K., B. JONES, C. HAVASI, H. LIEBERMAN y R. PICARD (2012), «Common sense reasoning for detection, prevention, and mitigation of cyberbullying», *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **2**(3), pág. 18.

DINAKAR, K., R. REICHART y H. LIEBERMAN (2011), «Modelling the Detection of Textual Cyberbullying.», en *The Social Mobile Web*, tomo 11, págs. 11–17.

DNEGRI, C. E. y E. L. DE VITO (2006), «Introducción al razonamiento aproximado: lógica difusa», *Revista Argentina de medicina respiratoria*, **4**, págs. 126–136.

DODDS, P. S. y C. M. DANFORTH (2010), «Measuring the happiness of large-scale written expression: Songs, blogs, and presidents», *Journal of Happiness Studies*, **11**(4), págs. 441–456.

ERIKSSON, G. y J. KARLGREN (2012), «Features for modelling characteristics of conversations: Notebook for PAN at CLEF 2012», en *CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, 17-20 September 2012*.

ESPELAGE, D. L. y S. M. SWEARER (2003), «Research on school bullying and victimization: What have we learned and where do we go from here?», *School Psychology Review*, **32**(3), págs. 365–384.

FAYYAD, U. M., G. PIATETSKY-SHAPIRO, P. SMYTH y R. UTHURUSAMY (1996), «Advances in knowledge discovery and data mining», *MIT Press*.

FELDMAN, R. (2013), «Techniques and Applications for Sentiment Analysis», Communications of the ACM, Vol. 56 No. 4, Pages 82-89, URL `http://cacm.acm.org/magazines/2013/4/162501-techniques-and-applications-for-sentiment-analysis/fulltext`.

FELDMAN, R. y I. DAGAN (1995), «Knowledge Discovery in Textual Databases (KDT).», en *KDD*, tomo 95, págs. 112–117.

FELDMAN, R. y J. SANGER (2007), *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press.

FRANKLIN, C. (2003), «SpectorSoft simplifies snooping: Every word they, type, every link they click, SpectorSoft Professional and eBlaster 3. O will be watching.», *InfoWorld*, (19), págs. 28–29.

GALVEZ, C. (2012), «Identifying and annotating generic drug names», en IEEE (editor), *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on*, págs. 1–6.

González Morcillo, C. (2012), *Lógica difusa, Una Introducción práctica*, Escuela Superior de Informatica - Universidad de la Mancha.

Goyal, P. y G. S. Kalra (2013), «Peer-to-Peer Insult Detection in Online Communities», *IIT*.

Harris, S. y G. Petrie (2002), «A study of bullying in the middle school», *NASSP Bulletin*, **86**(633), págs. 42–53.

Hidalgo, J. M. G. y A. A. C. Díaz (2012), «Combining Predation Heuristics and Chat-Like Features in Sexual Predator Identification.», en *CLEF (Online Working Notes/Labs/Workshop)*.

Hinduja, S. y J. W. Patchin (2008), *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*, Sage.

Hotho, A., A. Nurnberger y G. Paass (2005), «A Brief Survey of Text Mining.», en *Ldv Forum*, tomo 20, págs. 19–62.

Inches, G. y F. Crestani (2012), «Overview of the International Sexual Predator Identification Competition at PAN-2012», en *CLEF (Online Working Notes/Labs/Workshop)*, tomo 30.

Juvonen, J. y E. F. Gross (2008), «Extending the school grounds?—Bullying experiences in cyberspace», *Journal of School health*, **78**(9), págs. 496–505.

K Jowalski R, A. P., Limber S (2010), «Cyberbullying, el acoso escolar en la er@ digit@l.», en . Ed. Desclée Ce Brower (editor), *Cyberbullying, el acoso escolar en la er@ digit@l.*

Kern, R., S. Klampfl y M. Zechner (2012), «Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification», en *CLEF (Online Working Notes/Labs/Workshop)*.

Kiesler, S. (2014), *Culture of the Internet*, tomo 1, Psychology Press.

Kleinberg, J. M. (1999), «Authoritative sources in a hyperlinked environment», *Journal of the ACM*, **46**(5), págs. 604–632.

Klir, G. J., U. St Clair y B. Yuan (1997), *Fuzzy set theory: foundations and applications*, Prentice-Hall, Inc.

Kontostathis, A. (2009), «Chatcoder: Toward the tracking and categorization of internet predators», en SPARKS (editor), *Proc. Text Mining Workshop 2009 Held in conjunction with the ninth SIAM International Conference on data mining (SDM 2009)*, Citeseet, págs.–.

Kontostathis, A., L. Edwards y A. Leatherman (2010), «Text mining and cybercrime», *Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK*.

Kontostathis, A., A. Garron, K. Reynolds, W. West y L. Edwards (2012), «Identifying Predators Using ChatCoder 2.0.», en *CLEF (Online Working Notes/Labs/Workshop)*.

Kowalski, R. M., G. W. Giumetti, A. N. Schroeder y M. R. Lattanner (2014), «Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth.», *Psychological bulletin*, **140**(4), pág. 1073.

Liu, B. (2007), *Web data mining: exploring hyperlinks, contents, and usage data*, Springer.

Liu, B. (2010), «Sentiment analysis and subjectivity», *Handbook of natural language processing*, **2**, pág. 568.

Liu, B. (2012), «Sentiment analysis and opinion mining», *Synthesis Lectures on Human Language Technologies*, **5**(1), págs. 1–167.

Liu, H., H. Lieberman y T. Selker (2003), «A model of textual affect sensing using real-world knowledge», en *Proceedings of the 8th international conference on Intelligent user interfaces*, ACM, págs. 125–132.

López Lucio, L. A. (2009), «El cyberbullying en estudiantes de educación superior en México», *Congreso Nacional de Investigación Educativa*, **17**.

Macbeth, J., H. Adeyema, H. Lieberman y C. Fry (2013), «Script-based story matching for cyberbullying prevention», en *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, ACM, págs. 901–906.

Machmutow, K., S. Perren, F. Sticca y F. D. Alsaker (2012), «Peer victimisation and depressive symptoms: can specific coping strategies buffer the negative impact of cybervictimisation?», *Emotional and Behavioural Difficulties*, **17**(3-4), págs. 403–420.

Maher, D. *et al.* (2008), «Cyberbullying: An ethnographic case study of one Australian upper primary school class», *Youth Studies Australia*, **27**(4), pág. 50.

Manning, C. D., P. Raghavan, H. Schutze *et al.* (2008), *Introduction to information retrieval*, tomo 1, Cambridge university press Cambridge.

McGhee, I., J. Bayzick, A. Kontostathis, L. Edwards, A. McBride y E. Jakubowski (2011), «Learning to identify Internet sexual predation», *International Journal of Electronic Commerce*, **15**(3), págs. 103–122.

McNeill, D. y P. Freiberger (1994), *Fuzzy logic: The revolutionary computer technology that is changing our world*, Simon and Schuster.

Mendel, J. M. (2001), *Uncertain rule-based fuzzy logic system: introduction and new directions*, Prentice–Hall PTR.

Mendoza, L. (2012), «Acoso cibernético o cyberbullying: Acoso con la tecnología electrónica», *CIBERPEDS-CONAPEME*, **14**(3), págs. 133–146.

M.F, M. N. (2002), *Hermes: Servidor y biblioteca de modelos de recuperación de información*, Escuela de Ingenería, Universidad de las Américas Puebla.

MORRIS, C. y G. HIRST (2012), «Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features», en *CLEF (Online Working Notes/Labs/Workshop)*.

NADALI, S., M. A. A. MURAD, N. M. SHAREF, A. MUSTAPHA y S. SHOJAEE (2013), «A review of cyberbullying detection: An overview», en *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on*, IEEE, págs. 325–330.

NAHAR, V., L. XUE y P. CHAOYI (2012), «An Effective Approach for Cyberbullying Detection», *Communications in Information Science and Management Engineering*.

NARAYANAN, R., B. LIU y A. CHOUDHARY (2009), «Sentiment analysis of conditional sentences», en *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics, págs. 180–189.

NAVARRO, R., C. SERNA, V. MARTÍNEZ y R. RUIZ-OLIVA (2012), «The role of Internet use and parental mediation on cyberbullying victimization among Spanish children from rural public schools», *European Journal of Psychology of Education*, págs. 1–21.

NCMEC (2008), «National Center of missiong and exploited children», URL `http://www.missingkids.com/home`.

NOSWEARING.COM (), «Bad Word List & Swear Filter», URL `http://www.noswearing.com/`.

OLAH, L., F. FRIEDLER y Z. KOVACS (2002), «System and method for monitoring computer usage», .

PANG, B. y L. LEE (2004), «A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts», en *Proceedings of the 42nd*

annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pág. 271.

PANG, B. y L. LEE (2008), «Opinion mining and sentiment analysis», *Foundations and trends in information retrieval*, **2**(1-2), págs. 1–135.

PANG, B., L. LEE y S. VAITHYANATHAN (2002), «Thumbs up?: sentiment classification using machine learning techniques», en *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Association for Computational Linguistics, págs. 79–86.

PARAPAR, J., D. E. LOSADA y A. BARREIRO (2012), «A Learning-Based Approach for the Identification of Sexual Predators in Chat Logs», en *CLEF (Online Working Notes/Labs/Workshop)*.

PATCHIN, J. W. y S. HINDUJA (2006), «Bullies move beyond the schoolyard a preliminary look at cyberbullying», *Youth violence and juvenile justice*, **4**(2), págs. 148–169.

PEERSMAN, C., F. VAASSEN, V. VAN ASCH y W. DAELEMANS (2012), «Conversation Level Constraints on Pedophile Detection in Chat Rooms.», en *CLEF (Online Working Notes/Labs/Workshop)*.

PIDGIN (2015), «Pidgin community», URL `http://pidgin.im/`.

POPESCU, M. y C. GROZEA (2012), «Kernel Methods and String Kernels for Authorship Analysis», en *CLEF (Online Working Notes/Labs/Workshop)*.

POTHA, N., M. MARAGOUDAKIS y D. LYRAS (2016), «A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data», *Knowledge-Based Systems*.

PTASZYNSKI, M., P. DYBALA, T. MATSUBA, F. MASUI, R. RZEPKA y K. ARAKI (2010), «Machine learning and affect analysis against cyber-bullying», *the 36th AISB*, págs. 7–16.

Rada Vilela, J. (2013), «Software qtfuzzylite, aplicación de lógica difusa.», en *Fuzzylite: a fuzzy logic control*.

Reynolds, K., A. Kontostathis y L. Edwards (2011), «Using machine learning to detect cyberbullying», en *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, tomo 2, IEEE, págs. 241–244.

Salmivalli, C., K. Lagerspetz, K. Bjorkqvist, K. Osterman y A. Kaukiainen (1996), «Bullying as a group process: Participant roles and their relations to social status within the group», *Aggressive behavior*, **22**(1), págs. 1–15.

Sameer, H. (2015), «Cyberbullying Research Center», en *Cyberbullying Research Center*, URL `http://cyberbullying.org/`.

Sameer, H. y J. Patchin (2008), «Cyberbullying: An exploratory analysis of factors related to offending and victimization.», *Deviant Behavior*, **29**, págs. 129–156.

Sánchez, C. S. P. y. V. O. S. y. D. M. A., Jimena López y Carolina (2013), «Cuéntaselo a quien más confianza le tengas...? Facebook o Twitter?», *Centro Universitario Anglo Mexicano*.

Sanchez, H. y S. Kumar (2011), «Twitter bullying detection», *UCSC ISM245 Data Mining course report*.

Ševčíková, A. y D. Šmahel (2009), «Online harassment and cyberbullying in the Czech Republic: Comparison across age groups», *Zeitschrift für Psychologie/Journal of Psychology*, **217**(4), págs. 227–229.

Simanjuntak, D. A., H. P. Ipung, C. Lim y A. S. Nugroho (2010), «Text Classification Techniques Used to Faciliate Cyber Terrorism Investigation», en *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, IEEE, págs. 198–200.

Slonje, R., P. K. Smith y A. Frisén (2012), «The nature of cyberbullying, and strategies for prevention», *Computers in Human Behavior*.

SMITH, P. K. (1999), *The nature of school bullying: A cross-national perspective*, Psychology Press.

SMITH, P. K. y G. COLLAGE (2006), «Ciberacoso: naturaleza y extensión de un nuevo tipo de acoso dentro y fuera de la escuela», en *Congreso Educación Palma de Mallorca*.

SOOD, S., J. ANTIN y E. CHURCHILL (2012), «Using Crowdsourcing to Improve Profanity Detection», en *AAAI Spring Symposium Series*, págs. 69–74.

STICCA, F. y S. PERREN (2012), «Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying», *Journal of youth and adolescence*, págs. 1–12.

SUI, J. (2015), *Understanding and fighting bullying with machine learning*, Tesis Doctoral, University of Wisconsin.

SULER, J. (2004), «The online disinhibition effect», *Cyberpsychology & behavior*, **7**(3), págs. 321–326.

TABOADA, M., J. BROOKE, M. TOFILOSKI, K. VOLL y M. STEDE (2011), «Lexicon-based methods for sentiment analysis», *Computational linguistics*, **37**(2), págs. 267–307.

TAN, A.-H. *et al.* (1999), «Text mining: The state of the art and the challenges», en *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, tomo 8, pág. 65.

TAN, P.-N., F. CHEN y A. JAIN (2010), «Information assurance: Detection of web spam attacks in social media», en *Proceedings of Army Science Conference, Orland, Florida*.

TARAPDAR, S. y M. KELLETT (2012), «Cyberbullying: Insights and Age-Comparison Indicators from a Youth-Led Study in England», *Child Indicators Research*, págs. 1–17.

Tokunaga, R. S. (2010), «Following you home from school: A critical review and synthesis of research on cyberbullying victimization», *Computers in Human Behavior*, **26**(3), págs. 277–287.

Tolosa, G. H. y F. R. Bordignon (2008), «Introducción a la Recuperación de Información», en e prints (editor), *Introducción a la Recuperación de Información*, *1*, tomo 1, Tolosa y Bordignon, pág. 149.

Tsur, O., D. Davidov y A. Rappoport (2010), «ICWSM-A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.», en *ICWSM*.

Turney, P. D. (2002), «Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews», en *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, págs. 417–424.

Vairinhos, V. M. (2003), «Desarrollo de un sistema para minería de datos basado en los métodos Biplot», .

Vartapetiance, A. y L. Gillam (2012), «Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification», en *Proceedings of the 6th PAN workshop at CLEF2012 on uncovering plagiarism, authorship, and social software misuse (PAN2012), Rome*, Citeseer.

Ventura de Souza, L. (2011), «Análise de sentimentos no Twitter utilizando SentiWordNet», Proposta de Trabalho de Graduação.

Villatoro-Tello, E., A. Juárez-González, H. J. Escalante, M. Montes-y Gómez y L. V. Pineda (2012), «A Two-step Approach for Effective Detection of Misbehaving Users in Chats.», en *CLEF (Online Working Notes/Labs/Workshop)*.

Walrave, M. y W. Heirman (2011), «Cyberbullying: Predicting victimisation and perpetration», *Children & Society*, **25**(1), págs. 59–72.

WEBSTER, R. (2013), «Ciberacoso y Suicidios, http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html», URL `http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html`.

WELLMAN, B. (1999), «Networks in the global village», JSTOR.

WILLARD, N. (2007), *Cyberbullying and Cyberthreats (Book and CD): Responding to the Challenge of Online Social Aggression, Threats, and Distress*, Research press.

WONG, S. K. M. y Y. YAO (1992), «An information-theoretic measure of term specificity», *Journal of the American Society for Information Science*, **43**(1), págs. 54–61.

XU, J.-M., K.-S. JUN, X. ZHU y A. BELLMORE (2012), «Learning from bullying traces in social media», en *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, Association for Computational Linguistics, págs. 656–666.

KEY: xu2012learning

ANNOTATION: @inproceedingsxu2012learning, title=Learning from bullying traces in social media, author=Xu, Jun-Ming and Jun, Kwang-Sung and Zhu, Xiaojin and Bellmore, Amy, booktitle=Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies, pages=656–666, year=2012, organization=Association for Computational Linguistics

YANG, J. y Q. L. Y. ZHUANG (2015), «Modeling Data and User Characteristics by Peer Indexing», en *Tamkang University, Taiwan*, Citeseer.

YIN, D., Z. XUE, L. HONG, B. D. DAVISON, A. KONTOSTATHIS y L. EDWARDS (2009), «Detection of harassment on web 2.0», *Proceedings of the Content Analysis in the WEB*, **2**, págs. 1–7.

YU, H. y V. HATZIVASSILOGLOU (2003), «Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences», en *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, págs. 129–136.

ZADEH, L. A. (1996), «Fuzzy logic= computing with words», *Fuzzy Systems, IEEE Transactions on*, **4**(2), págs. 103–111.

Resumen autobiográfico

Laura Patricia Del Bosque Vega

Candidato para obtener el grado de
Doctorado en Ingeniería
con Orientación en Tecnologías de la Información

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica

Tesis:
Detección automática de ciberacoso en redes sociales

Soy Laura Patricia Del Bosque Vega, nací un 17 de Febrero de 1982, en la ciudad de Monclova Coahuila, estoy casada desde hace 9 maravillosos años, mi esposo es el Lic. Mario Eduardo Carrillo Ruiz, mis padres son Lic. Jorge Humberto Del Bosque Canseco y Sra. Laura Leticia Vega Aguilera, mis hermanos Jorge Antonio y Adriana, mi cuñada Magdely y mi sobrino Koku.

Soy Ingeniero Administrador de Sistemas, orgullosamente egresada de la Facultad de Ingenieriá Mecánica y Eléctrica (FIME) de la Universidad Autonoma de Nuevo Leon (UANL), en el ańo del 2003, a mucha honra Tigre y Oso.

Estudie la maestria en ingeniería con Orientación en Informatica, terminandola en Mayo del 2009, en FIME.

He trabajado en diferentes empresas desde 5to semestre de la licenciatura, por ejemplo en consultorías de desarrollo de software; en empresas como DALTILE, Corporativo de Soriana y en el ámbito educativo: Preparatoria INSUCO, la cual fungí como directora y en la actualidad soy la Jefe del programa educativo de Ingeniero Administrador de Sistemas desde el año 2013 a la fecha en la FIME.

A su vez pertenezco como evaluador del Consejo de Acreditación de la Enseñanza de la Ingenieria, A.C., miembro de consejos consultivos para programas educativos tanto de preparatorias técnicas como de licenciatura de la UANL.