EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Detección de apología del delito en redes sociales utilizando un modelo difuso

presentada por

Ing. Araceli Alvarez Cruz

como requisito para la obtención del grado de

Maestra en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Héctor Jiménez Salazar

Cuernavaca, Morelos, México. Julio de 2022.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 10/junio/2022

OFICIO No. DCC/043/2022
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de la C. ARACELI ÁLVAREZ CRUZ, con número de control M18CE079, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"DETECCIÓN DE APOLOGÍA DEL DELITO EN REDES SOCIALES UTILIZANDO UN MODELO DIFUSO"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ
Director de tesis

DR. HÉCTOR JIMÉNEZ SALAZAR
Codirector de tesis

DR. DANTE MÚJICA VARGAS
Revisor 1

DR. JUAN GABRIEL GONZÁLEZ SERNA
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante
JGGS/ibm

Cuernavaca, Mor., 30/junio/2022
No. De Oficio: SAC/111-2/2022
Asunto: Autorización de impresión de tesis

ARACELI ÁLVAREZ CRUZ
CANDIDATO(A) AL GRADO DE MAESTRO(A) EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"DETECCIÓN DE APOLOGÍA DEL DELITO EN REDES SOCIALES UTILIZANDO UN MODELO DIFUSO"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Educación Tecnológica al Servicio de México"

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/CHG

Dedicatoria

El presente trabajo de investigación lo dedico principalmente a Dios, por darme día a día la fuerza necesaria para continuar en este proceso y poder cumplir con uno de mis anhelos más deseados.

A mis padres Indalecio y Beneda; por su amor, trabajo y sacrificio en todos estos años, gracias por inculcar en mí el ejemplo de esfuerzo y perseverancia ante las distintas adversidades que se han presentado en mi camino; por estar conmigo en todo momento y brindarme palabras de aliento cuando más las necesito, les agradezco por su cariño y apoyo incondicional durante todo este proceso porque gracias a ellos he podido llegar hasta aquí y convertirme en lo que soy. De verdad gracias por estar siempre conmigo y por mostrarme su amor incondicionalmente.

A mi hermana Graciela que en paz descansé; su recuerdo siempre ha sido mi motor e inspiración para seguir adelante; gracias a ella entendí que debo ser tenas y perseverante en mis metas a pesar de estás no sean del todo aceptadas por los demás.

A mi pareja Gabriel porque, aunque no me dijera las palabras que quería escuchar estuvo ahí para apoyarme durante el tiempo que duro todo este proceso a pesar de que en ocasiones dudaba que lo fuera lograr y en algunas ocasiones hizo sentirme incapaz de lograrlo gracias a todo eso tuve la fuerza y tenacidad suficiente para poder culminar con esta etapa que me hace muy feliz.

Finalmente, pero no por eso menos importante quiero dedicar este trabajo a mi hijo Gabriel, el cual llego a mi vida para enseñarme que los cambios no son tan malos o difíciles como lo pensaba, solo hay que adaptarse a las situaciones que se nos presentan y dar lo mejor de nosotros; gracias por enseñarme a ser una mejor versión de mi cada día con el fin de hacerte sentir orgullo de mí.

Agradecimientos

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado durante el desarrollo de esta tesis para la obtención del grado de Maestra en Ciencias de la Computación mediante su sistema de becas de posgrado, el cual me permitió desempeñarme como estudiante de posgrado a tiempo completo.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), perteneciente al (TECNM), por brindarme la oportunidad de superarme académicamente mediante el programa de Maestría en Ciencias y por prestarme sus instalaciones para poder cumplir con dicho objetivo.

Quiero agradecer principalmente a mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, por la oportunidad y confianza brindada para desarrollar está investigación; gracias por sus consejos, paciencia y tiempo brindado durante mi formación como maestra.

A mi codirector, el Dr. Héctor Jiménez Salazar por cada una de sus observaciones, consejos y sugerencias a lo largo de esta tesis.

A mis revisores, el Dr. Juan Gabriel González Serna y el Dr. Dante Mújica Vargas, quienes dedicaron parte de su tiempo a realizar las revisiones necesarias para fortalecer este trabajo de investigación, gracias por el apoyo y confianza brindada.

Por último, quiero agradecer a todos aquellos docentes que directa o indirectamente participaron en mi formación académica durante el desarrollo de la maestría.

Resumen

Las redes sociales en la actualidad son uno de los principales medios de comunicación que permiten a los usuarios interactuar entre un mismo grupo de personas con intereses en común o diferentes gustos compartiendo opiniones sobre temas relacionados con política, artes, diseño, entretenimiento, ciencia, etc. Este tipo de redes permiten a los usuarios realizar intercambio de información mediante videos, audios, mensajes, imágenes entre otros más.

De igual manera mediante el uso de las redes sociales, los usuarios pueden ejercer uno de sus principales derechos como ciudadanos, el cuál es la libre expresión u opinión sobre algún tema en este especifico. Sin embargo, últimamente se ha observado que en dichas plataformas el contenido que se comparte presenta algún tipo motivación o invitación para cometer acciones delictivas afectando psicológica y moralmente a los usuarios; este tipo de invitaciones son consideradas como apologías de delito.

En esta tesis se trabajó con la red social de Twitter a fin de poder analizar la información que comparten sus usuarios mediante la publicación de comentarios, mediante un modelo difuso que asigna grados de pertenencias y los clasifica como aquellos que presentan contenido apológico y aquellos que no lo presentan.

Adicionalmente se muestra la creación de un corpus con comentarios extraídos de Twitter, que muestran contenido que incita a sus usuarios a cometer algún tipo de apología del delito; otra actividad que se realizó fue la creación de la base de conocimiento que conforma al modelo difuso, la cual se compone por reglas difusas y las funciones de pertenencia. El modelo difuso creado está basado en el modelo de Takagi Sugeno, con el propósito de clasificar mediante inferencia difusa comentarios en las categorías de Apológico y No Apológico, dependiendo el grado de pertenencia que se presenta en el análisis del contenido del comentario.

Para evaluar el modelo difuso se utilizaron las métricas de Precisión, Cobertura y F1-Store.

Abstract

Social networks today are one of the main means of communication that allow users to interact between the same group of people with common interests or different tastes, sharing opinions on topics related to politics, arts, design, entertainment, science, etc. This type of network allows users to exchange information through videos, audios, messages, images, among others.

In the same way, through the use of social networks, users can exercise one of their main rights as citizens, which is free expression or opinion on a specific topic. However, lately it has been observed that on these platforms the content that is shared presents some type of motivation or invitation to commit criminal actions affecting users psychologically and morally; this type of invitations are considered as apologies of crime.

In this thesis, we worked with the Twitter social network in order to be able to analyze the information shared by its users through the publication of comments, through a fuzzy model that assigns degrees of belonging and classifies them as those that present apologetic content and those that do not. they present it.

Additionally, the creation of a corpus with comments extracted from Twitter is shown, which show content that encourages its users to commit some type of defense of the crime; Another activity that was carried out was the creation of the knowledge base that conforms to the fuzzy model, which is made up of fuzzy rules and membership functions. The fuzzy model created is based on the Takagi Sugeno model, with the purpose of classifying comments into the Apological and Non-Apologetic categories by means of diffuse inference, depending on the degree of belonging that is presented in the analysis of the content of the comment.

To evaluate the fuzzy model, the Accuracy, Coverage and F1-Store metrics were used.

Índice

Lista de Figuras

Lista de Tablas

Capítulo 1 Introducción

Actualmente las redes sociales son consideradas como aplicaciones que permiten la comunicación entre individuos, grupos de personas y/o organizaciones que presentan intereses en común, con el fin de obtener una fácil interacción entre sí mismos para compartir información, datos y contenidos de diferentes formatos.

Dichas plataformas sociales son el medio de comunicación que permiten a los usuarios ejercer su derecho de libertad de expresión, través de ellas se convocan eventos multitudinarios como lo son invitaciones a conciertos musicales, conferencias, eventos políticos, eventos deportivos y manifestaciones.

Asimismo, se ha observado en ellas que la presencia de comentarios que motivan o invitan a cometer actos o hechos constitutivos de delitos contra una persona o grupo de personas ha ido incrementando. En dichos comentarios se observan acciones que motivan a cometer acciones delictivas como incitaciones al odio o violencia, calumnias, amenazas, acosos, delitos contra la intimidad, injurias y humillaciones; este tipo de acciones también son conocidas como apologías del delito *(Torrús, 2018).*

A nivel mundial México es considerado como el segundo país de América Latina con mayor porcentaje de presencia de usuarios que usan la red social Twitter; dicha red es considerada como la sexta plataforma social que predomina en México contando con 14 millones de cuentas de las cuales el 56% de ellas se encuentran activas *(Statistac, 2022).*

Esta plataforma social es un servicio que permite a sus usuarios comunicar noticias y aportar opiniones sobre algún tema en particular mediante mensajes rápidos y frecuentes, los cuales son conocidos como Tweets y pueden estar compuestos por fotos, videos, enlaces y textos. El desarrollo de dicha plataforma social permite llevar a cabo el estudio de fenómenos sociales que se presentan en la actualidad, en los cuales se ven involucradas diversas formas de comunicación y expresión, dirigidas contra individuos o grupos por su religión, raza, orientación sexual, discapacidad, etnia, nacionalidad, entre otras.

En la actualidad se puede observar que la presencia de conductas apológicas está presente en varias plataformas sociales y aunque existen algoritmos de aprendizaje automático enfocados a la identificación de dichas conductas, estas no han sido censuradas en su totalidad; debido que la ambigüedad del lenguaje es bastante extensa por lo que carecen de la capacidad necesaria para llevar a cabo la diferencia entre el lenguaje formal y el lenguaje coloquial.

En este trabajo de investigación se estará abordando el tema de tesis “Detección de Apología del Delito en redes sociales utilizando un modelo difuso”; el cual tiene como objetivo principal desarrollar un método que utilice Procesamiento de Lenguaje Natural y Lógica Difusa para analizar e identificar en la red social Twitter que contengan expresiones que hagan referencia a apologías del delito.

1.1 Planteamiento del problema

En los últimos años se ha observado un incremento significativo en el uso de las redes sociales a nivel mundial. En México, el número de usuarios de redes sociales se mantiene en constante aumento, y se prevé que supere los 90 millones de usuarios en 2023 *(Statistac, 2022).*

Entre los diversos usos que se le dan a las redes sociales, se han detectado comentarios sobre apología del delito que generan alarma y desconcierto entre los usuarios; y por otro lado logran su cometido al ser considerados como una forma de provocación constituyendo una incitación directa a cometer un delito, entre los que destacan: acciones de odio, violencia contra grupos o asociaciones por motivos raciales, ideológicos o religiosos, provocación sexual, robo, extorsión, estafa, etc *(Jaconelli, 2018).*

Dado que actualmente en redes sociales las incitaciones de odio o violencia son actividades delictivas que comúnmente se presentan, se puede observar que miles de personas son atacadas de manera psicológica o físicamente cada año por delitos de odio, debido que tales ofensas no solo afectan a las víctimas, sino también los pensamientos y el comportamiento de los demás.

Durante los últimos cinco años, el Proyecto de Crimen de Odio en la Universidad de Sussex ha investigado estos impactos odio, observando cómo el simple hecho de conocer a una víctima o incluso de escuchar un incidente puede tener consecuencias significativas *(Brown, Walters, & Paterson, 2018).*

1.2 Justificación

Las redes sociales no solo son aplicaciones y portales usados por los usuarios para colocar fotos, buscar amigos, sino que también sirven para hacer política y actividades delictivas que perjudican a los usuarios que utilizan estas redes, por lo que es importante hacer uso de una herramienta que permita identificar cuáles son los mensajes que presentan incitaciones delictivas, las cuales pueden causar que los usuarios de estas redes sean partícipes en delitos o víctimas de internautas que utilizan estas plataformas para delinquir.

En México se han desarrollado unidades de policía cibernética, entre las cuales destacan la Policía de Ciberdelincuencia Preventiva de la Ciudad de México, la división científica y el centro nacional de respuesta a incidentes cibernéticos (CERT-MX) de la Policía Federal (PF) cuya labor reside en detectar y prevenir amenazas. Sin embargo la falta de credibilidad en la preparación técnica de las autoridades para resolver estos delitos es uno de los factores que desincentiva a la denuncia *(Universal, 2016)*.

Con este trabajo se propone un método que analice comentarios en redes sociales para identificar expresiones que inciten a cometer delitos de odio o violencia, a través del monitoreo y la extracción automática de comentarios, y su procesamiento mediante un modelo de lógica difusa.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un método que utilice Procesamiento de Lenguaje Natural y Lógica Difusa para analizar e identificar comentarios en la red social Twitter que contengan expresiones que hagan referencia a apología del delito.

1.3.2 Objetivos específicos

- Obtener o generar un corpus de comentarios de la red social Twitter que hagan apología del delito.
- Conocer el funcionamiento de la lógica difusa para su implementación.
- Definir las características lingüísticas utilizadas en mensajes que hacen apología del delito.
- Adquirir comentarios de la red social Twitter mediante el desarrollo de un algoritmo automático de descarga.
- Implementar el modelo de inferencia difusa para la clasificación de apología de delito.

1.4 Alcances y limitaciones

1.4.1 Alcances

- El modelo procesará comentarios publicados en la red social Twitter.
- Se generará un corpus compuesto por textos apológicos los cuales estarán disponibles para ser usados por otros investigadores.

1.4.2 Limitaciones

- La herramienta que se desarrollará solo trabajará con el idioma español de México.
- La herramienta que se desarrollará solo se enfocará en apologías del delito como actividades de odio o violencia.

1.5 Organización del documento

Este documento presenta una estructura de seis capítulos que integran la definición de argumentos, investigación, desarrollo, evaluación, resultados y conclusión de tema de tesis. A continuación, se da una breve explicación sobre el contenido de cada capítulo.

Capítulo 1. Introducción: En ese capítulo se describe el origen de la investigación, el problema que se trata de solucionar y la justificación para elaborar esta tesis, así como sus objetivos, alcances y limitaciones.

Capítulo 2. Marco Conceptual: En este capítulo se describen algunos de los conceptos utilizados en el desarrollo de este trabajo de tesis, los cuales son indispensables para para tener una mejor compresión al momento de dar lectura al presente trabajo.

Capítulo 3. Estado del Arte: En este capítulo se presenta un análisis de los trabajos relacionados con esta tesis, los cuales se clasifican en dos categorías, Análisis de violencia en texto y Lógica difusa.

Capítulo 4. Método Propuesto: En este capítulo se describe a detalle el método propuesto para identificar comentarios con contenido apológico; el cual se compone por cuatro fases: Creación del Corpus, Procesado de Comentarios, Definición de Reglas las cuales son importantes para la creación del modelo Difuso y Evaluación del modelo.

Capítulo 5. Pruebas y Resultados: En este capítulo se muestran las diferentes pruebas realizadas con el modelo difuso. Así como los resultados obtenidos de dicho modelo considerando las métricas de evaluación.

Capítulo 6. Conclusiones: En este capítulo se presentan las conclusiones derivadas de la investigación, así como algunas propuestas para trabajos futuros.

Capítulo 2 Marco Conceptual

En esta sección se presentan algunos conceptos utilizados en este trabajo de investigación para una óptima comprensión del resto del documento.

2.1 Delitos informáticos

Corresponden a aquellas infracciones penales cometidas utilizando un medio o instrumento informático, donde cada vez más delincuentes aprovechan la velocidad, la comodidad y el anonimato de Internet para cometer una amplia gama de actividades delictivas que no conocen fronteras, ya sean físicas o virtuales, causan graves daños y representan amenazas muy reales para las víctimas en todo el mundo *(Interpol, 2018).*

2.2 Apología del delito

Apología proviene del latín apología y significa “discurso en defensa o alabanza de persona o cosa” y delito proviene del latín delicto y significa “culpa, crimen o quebrantamiento de la ley” por lo que el significado en su conjunto es de: alabanza de un quebrantamiento grave de la ley instigando de manera indirecta a la participación en una conducta delictiva *(Campuzano, 2018).*

Figura 1: Comentarios apológicos en redes sociales (Twitter, 2021)

2.3 Redes Sociales

Son estructuras sociales capaces de comunicar entre sí a personas, organizaciones y otras entidades. Los miembros de estas entidades a través de internet interactúan y crean entre ellos una comunidad virtual que persigue un relativo interés en común, estableciendo relaciones donde pueden compartir hobbies, religión, preferencias políticas, estilo de vida, etc *(Facchin, 2018).*

En Tabla 1, se muestra un conjunto de las principales redes sociales utilizadas actualmente por los usuarios, mostrando sus funciones, ventajas, desventajas y la cantidad de usuarios activos por mes.

Tabla 1:Tabla comparativa de redes sociales

Red Social	Descripción	Funciones/Objetivo	Ventajas/Beneficios	Desventajas	Usuarios activos por mes
facebook	Herramienta social que pone en contacto a personas con sus amigos y otras personas que trabajan y viven cerca de ellos o no.	Publicar información personal y profesional, fotos, chatear y ser parte de grupos según intereses personal.	• Método de ocio • Información en tiempo real. • Targeting (permite al anunciante dirigirse a un determinado tipo de público en función de sus intereses.	• Inseguridad al ingresar datos personales. • La privacidad es limitada.	Aproximadamente 2, 196 millones de usuarios alrededor del mundo.
twitter	Es un servicio de microblogging que promueve la interacción entre usuarios enviando micro-entradas basadas en textos, denominadas tweets.	Comentar y compartir acontecimientos cotidianos y que están pasando en cualquier parte del mundo.	• Método de ocio • Es relativamente editable, posibilitando la personalización a gusto del usuario. • Información en tiempo real.	• Falta de herramientas para incluir al usuario común al sistema	Aproximadamente 336 millones de usuarios
YouTube	Es un sitio web donde los usuarios pueden subir y compartir video.	Alojar videos personales de manera sencilla.	• Método de ocio. • Permite cargar varios videos al mismo tiempo. • Se puede interactuar con otros usuarios por medio de video o mensajes privados a tu canal	• Amenaza con la privacidad de las personas. • Los comentarios no tienen censura.	Aproximadamente 1, 900 millones de usuarios alrededor del mundo.
tumblr	Plataforma de microblogging que permite a usuarios publicar textos, imágenes, videos, citas y audio de manera que tumblelog.	Compartir fotos, videos, textos y citas en un solo lugar (dashboard).	• Se pueden subir archivos de multimedia y links de páginas. • Permite la interacción con otros usuarios por medio de chats.	• Constantemente se presentan problemas con el sistema de la página. • El uso de la plataforma es difícil debido a que está en inglés.	Aproximadamente 790 millones de usuarios.

myspace .com	Es una plataforma que permite al usuario subir información, comentarios, mensajes instantáneos y participar en foros de distintas orientaciones, gustos, ideologías y creencias.	Permite chatear, mandar mensajes, crear blogs, invitar amigos a participar, personalizar la página, subir fotos y videos.	• Método de ocio • Es otra forma de hacer amigos • Es veloz	• Te pueden robar las fotos no importa que estén protegidas. • Mucha cantidad de publicaciones	Aproximadamente 38 millones de usuarios.
Linked in	Sitio web orientado a grupos profesionales.	Ser utilizado por las empresas como un canal de comunicación y marketing.	• Se pueden generar oportunidades de trabajo. • Contacto directo con los líderes y referencias de industria. • Compartir información a través de grupos	• Se consiguen datos personales en cuestión de minutos. • Se puede difamar a una persona fácilmente, y ser vistos en poco tiempo por una multitud,	Aproximadamente 294 millones de usuarios alrededor del mundo.
Instagram	Es una red social que está orientada a fotografías realizadas en un momento en específico para ser compartidas en la red.	Ser utilizada por empresas para fortalecer su marca (branding) y acercarla a la audiencia.	• Permite ver y compartir fotografías con personas de otros países. • Se pueden ver publicaciones de artistas y ver desde que parte del mundo se subió dicha publicación. • Alto componente social.	• Las actualizaciones son constantes. • Existen muchas cuentas falsa o usuarios que se hacen pasar por artistas. • Solo es accesible para cargar fotos a través del móvil y no a través de la web.	Aproximadamente 1,000 millones de usuarios alrededor del mundo.
reddit	Es un sitio web donde los usuarios comparten enlaces a contenidos de otros sitios, como un agregador de noticias,	Ser utilizado como un sitio que contenga cualquier temática para que el usuario pueda acceder a la que desee.	• Permite realizar comentarios de publicaciones. • Realizar búsquedas sobre diferentes tipos de temáticas.	• Los usuarios son los que se encargan de convertir una publicación en popular mediante votos.	Aproximadamente 330 millones de usuarios

2.3.1 Misión de las redes sociales

La misión principal de las redes sociales es proporcionar a los usuarios la capacidad de expresar sus opiniones y creencias mediante la compartición de ideales, sin ningún tipo de obstáculos.

2.4 Clasificación, categorización y etiquetado

Dado un grupo de objetos, la tarea de clasificarlos consiste en asignarlos a un conjunto pre especificado de categorías. Si estamos dentro del dominio de gestión documental, la tarea se la conoce como categorización de texto, y consiste en hallar uno o más tópicos en los que encajen los contenidos de los documentos; teniendo

como entrada un grupo de categorías (sujetos – temas) y un conjunto de documentos de texto. La categorización automática de documentos es una forma de clasificación de patrones, que se es necesaria para la gestión eficiente de sistemas de información de textos *(Hernández & Gómez, 2013).*

2.5 Algoritmo de agrupación

Estos algoritmos agrupan un conjunto de documentos en subconjuntos o categorías. Su objetivo es crear categorías cuyos documentos sean similares entre sí, pero no tengan similitud con los de otros grupos. En los algoritmos no supervisados no se requieren categorías predefinidas para agrupar la información, mientras que en los supervisados es necesaria una fase de entrenamiento que le ayuda al algoritmo a tomar mejores decisiones al clasificar información *(Manning & Raghavan, 2008).*

2.6 Lenguaje Natural

El lenguaje natural se refiere al lenguaje que utilizan los seres humanos para comunicarse, también se puede definir como un conjunto de símbolos gráficos, verbales o gesticulares que se combinan para transmitir información de un individuo a otro *(Chopra, Prashar, & Sain, 2013).*

2.7 Procesamiento de Lenguaje Natural (PLN)

El procesamiento de lenguaje natural (PLN) es un subcampo de la inteligencia artificial (IA) y la lingüística, este subcampo se dedica a hacer que las computadoras comprendan el lenguaje humano *(Chopra, Prashar, & Sain, 2013).* La comprensión del lenguaje natural les permite a las computadoras realizar tareas como extracción de información, traducción automática, recuperación de información, resumen automático, etc.

2.8 Extracción de información

La extracción de información es una tarea de procesamiento del lenguaje natural, cuyo propósito es extraer determinados tipos de información de un documento. Los sistemas de EI (Extracción de Información) son específicos en un dominio, ya que extraen eventos o hechos particulares de un dominio concreto y omiten aquellos que no lo son *(García, 2004).*

2.9 Corpus

Un corpus se puede definir como una colección de textos, los cuales se convierten en repositorios de información a partir de los cuales se pueden encontrar, obtener, y, por lo tanto, aprender los múltiples contextos en los que puede aparecer una determinada palabra, convirtiéndose así, en una fuente de información fundamental para el sistema *(Taulé & Martí., 2003).*

2.10 Tokenización

Es el proceso de segmentar el texto en palabras y oraciones llamados tokens. El texto electrónico es una secuencia lineal de símbolos (caracteres o palabras o frases). Naturalmente, antes de que se realice cualquier procesamiento de texto real, el texto debe segmentarse en unidades lingüísticas tales como palabras, signos de puntuación, números, números alfanuméricos, etc. Este proceso se denomina tokenización. La tokenización es una especie de pre-procesamiento en cierto sentido; una identificación de las unidades básicas a procesar *(Trim, 2018).*

2.11 Stopwords

Las palabras vacías o stopwords se refieren a un conjunto de palabras que no poseen un significado o relevancia para el análisis de un texto y que se encuentran muy frecuentemente en todos los textos (artículos, pronombres, preposiciones, etc.) *(Mishra & Vishwakarma, 2015).* En el procesamiento de lenguaje natural, se recomienda utilizar listas de stopwords para filtrar los textos antes del análisis, con la finalidad de evitar interferencias en el resultado o demoras en el procesamiento.

2.12 Freeling

Freeling es una librería que proporciona servicios de análisis de lenguaje como: análisis morfológico, etiquetado PoS, clasificación de entidades nombradas, desambiguación semántica, detección de idioma, entre otros *(TALP Research Center, 2020).*

2.14 Lógica Difusa

La lógica difusa es una teoría de conjuntos propuesta por *(Zadeh, 1987).* En esta teoría se definen conjuntos difusos como una clase cuyos elementos cuentan con grados de pertenencia desde 0 a 1, entre más cercano sea el grado de pertenencia a 1 representa una mayor pertenencia y entre más cercana a 0 el caso contrario, *(Ponce-Cruz, Molina, & MacCleery, 2016).* Los conjuntos difusos tratan de modelar la incertidumbre relacionada al razonamiento natural humano, que se expresa con palabras y oraciones lingüísticas en lugar de expresiones matemáticas, el razonamiento difuso es un modo de razonamiento que no es exacto ni inexacto, para entenderlo, se deben comprender tres conceptos básicos:

- **Variable Lingüística.** Variable cuyos valores son palabras u oraciones en un lenguaje natural o artificial en lugar de numérico. Por ejemplo, la temperatura se puede describir como se presenta a continuación.

$$TS(temperatura)= \{Frío, Tibio, Caliente\} = \{F, T, C\}. \quad (2.1)$$

- **Proposición difusa**. Declaración expresada en un lenguaje natural o artificial. A diferencia de las proposiciones lógicas clásicas, puede adoptar un valor de verdad del intervalo [0, 1]. Por ejemplo, la temperatura es caliente.

- **Regla Lingüística.** Sentencia IF-THEN que se compone de dos partes: causa= IF {Proposición difusa}, consecuencia=THEN {Proposición difusa}.

Un conjunto difuso es caracterizado por la función de pertenencia 1, la cual asigna a cada elemento un grado de pertenencia. Un conjunto difuso *A* es definido por un conjunto de pares ordenados:

$$A = \{(x, \mu_A(x)) | x \in X\} \quad y \quad \mu_A \in [0,1] \tag{2.2}$$

Donde x es un elemento del universo U y μ_A es la función de pertenencia que se asigna a un grado de pertenencia $\mu_{\mathrm{A}}(x)$ para cada elemento x de *A (Mendel, 2001).*

Conjunto Difuso Tipo-1 (FST1)

Usualmente, el razonamiento humano en la toma de decisiones no es definido con métodos matemáticos, así que, los números difusos pueden ser usados para resolver problemas sencillos y avanzados que lidian con condiciones ambiguas. Los conjuntos difusos se utilizan para describir la falta de claridad en función de los grados de membresía y se pueden usar en muchas situaciones reales con términos lingüísticos *(Ponce-Cruz, Molina, & MacCleery, 2016).*

Función de Pertenencia

La función de pertenencia transforma cada elemento de X a un grado de pertenencia entre 0 y 1, *(Mendel, 2001).* Para entender mejor las funciones de pertenencia, se definen las siguientes nomenclaturas que se observan en la Figura 2 y se definen de la siguiente forma:

- **Soporte**: el soporte de un conjunto difuso *X*, es el conjunto de todos los puntos x en *X* tal que $\mu_{\mathrm{A}}(x) > 0$:

$$\mathrm{Soporte(A)} = \{x | \mu_{\mathrm{A}}(x) > 0\} \tag{2.3}$$

- **Núcleo**: es el conjunto de todos los puntos x en *X* tal que$\mu_{\mathrm{A}}(x) = 1$:

$$\mathrm{Núcleo(A)} = \{x | \mu_{\mathrm{A}}(x) = 1\} \tag{2.4}$$

- **Fronteras**: definen como aquellas regiones del universo que contienen elementos que tienen una pertenencia distinta de cero, pero no una pertenencia completa.

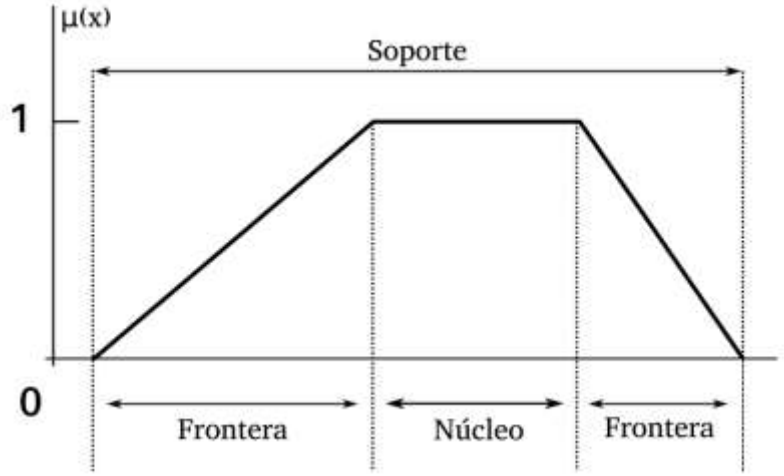
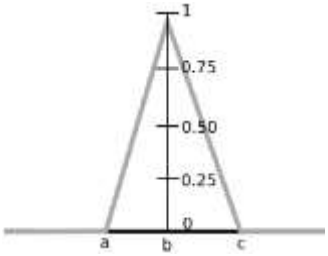


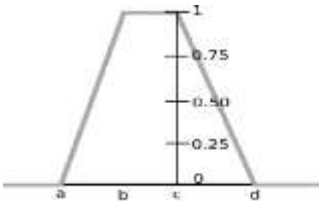
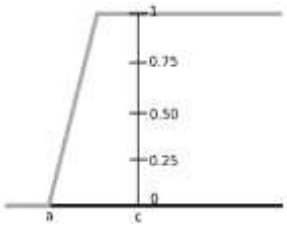
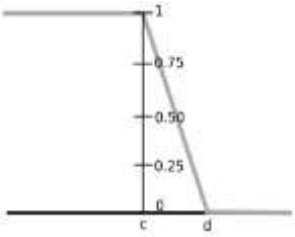
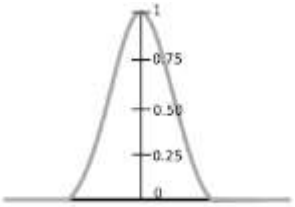
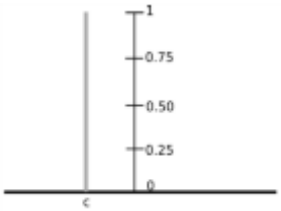
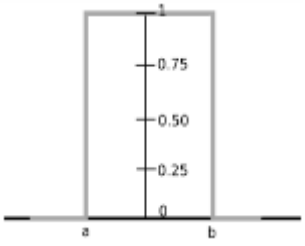
Figura 2: Partes de una función de pertenencia (Ponce-Cruz, Molina, & MacCleery, 2016).

Tipos de Funciones de Pertenencia

Existen varias funciones de pertenencia (MF) que se pueden usar en T1FS; por ejemplo, las funciones normales de membresía presentadas anteriormente. Un conjunto Fuzzy normal o convencional es aquel cuya MF tiene al menos un elemento x en el universo, cuyo valor de pertenencia, es uno. La representación matemática para algunas MFs convencionales se presentan en la Tabla 2.

Tabla 2: Funciones de pertenencia convencionales (Ponce-Cruz, Molina, & MacCleery, 2016).

Función de pertenencia	Parámetros	Forma
Triangular	$\mu_A(x) = \begin{cases} \frac{x-a}{b-a} & a \leq x \leq b \\ \frac{c-x}{c-b} & b \leq x \leq c \\ 0 & x \leq a \vee x \geq c \end{cases}$	

Trapezoidal	$\mu_A(x) = \begin{cases} \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \frac{d-x}{d-c} & c \le x \le d \\ 0\ x \le a \vee x \ge d \end{cases}$	
Forma-S	$\mu_A(x) = \begin{cases} \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \\ 0 & x \le a \end{cases}$	
Forma-Z	$\mu_A(x) = \begin{cases} 1 & x \le c \\ \frac{d-x}{d-c} & c \le x \le d \\ 0 & x \le d \end{cases}$	
Gaussiana	$\mu_A(x) = e^{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2}$	
Singleton	$\mu_A(x) = \begin{cases} 1 & x = c \\ 0 & otros \end{cases}$	
Intervalo	$\mu_A(x) = \begin{cases} 1 & a \le x \le b \\ 0 & otros \end{cases}$	

Sistema Lógico Difuso Tipo-1 (T1FLS)

Un sistema lógico difuso Tipo-1 basado en reglas contiene cuatro componentes principales: fuzzificación, base de conocimiento, inferencia y defuzzificación, que están interconectados como se muestra en la Figura 3 y se describen a continuación.

- **Fuzzificación:** convierte un valor de entrada discreto a un dominio difuso continuo por medio de una MF.
- **Base de conocimiento:** reglas canónicas conformadas por antecedentes y consecuentes de estructura IF-THEN; son establecidas de forma coherente por un experto; además, determina las formas, cantidad y umbrales de los conjuntos que se trabajaran para cada variable.
- **Inferencia:** este bloque asigna a la entrada difusa una salida difusa de acuerdo con las reglas establecidas y operaciones de conjuntos difusos.
- **Defuzzificación:** asignación de una entrada difusa de tipo-1 en una salida discreta que puede ser interpretada en forma de palabras (Por ejemplo: muy frio, tibio o demasiado caliente).

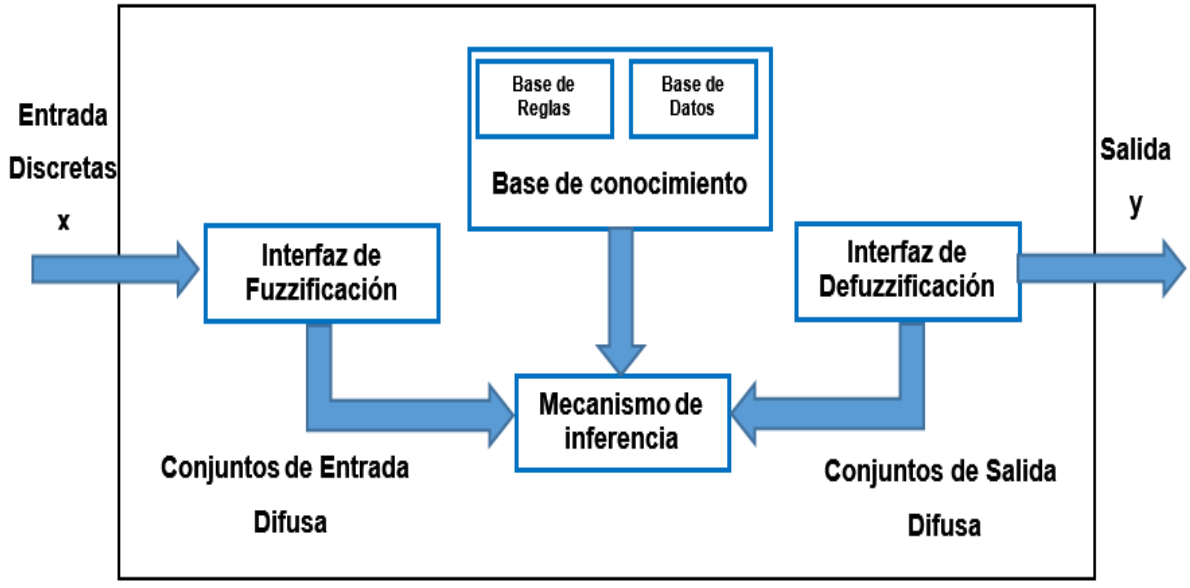


Figura 3:Diagrama de bloques de un sistema lógico difuso tipo-1 (Mendel, 2001).

Capítulo 3 Estado del Arte

En este capítulo se presentan los trabajos de investigación más representativos en la revisión de la literatura, relacionados con el tema de tesis.

3.1 Análisis de violencia en texto

Dentro del estudio de la literatura se identificaron algunos artículos que realizan el análisis del texto para identificar en ellos algún tipo de violencia, tal es el caso de los autores (Alghofaili & Almishari, 2018), los cuales analizan las cuentas de los usuarios a fin de poder identificar con precisión si una cuenta está promoviendo el extremismo religioso y la violencia; al igual que los autores *(Setyadi, Nasrun, & Setianingsih, 2018)* los cuales en su artículo de investigación mencionan la importancia de identificar palabras claves con relación a textos que invitan a cometer algún tipo de delito.

3.1.1 Creación de perfiles de Ciberdelito: técnicas de minería de textos para detectar y predecir actividades delictivas en publicaciones de microblogs *(Alami & Elbeqqali, 2015).*

En este trabajo de investigación se presentó una idea global sobre un sistema automático para detectar perfiles sospechosos en las redes sociales, a través del cual se logró descubrir comportamientos sospechosos e intereses de los usuarios, por lo que el enfoque propuesto se basa en el cálculo de una similitud.

El propósito del enfoque es descomponer cada publicación en términos, para compararlos automáticamente con la base de datos de términos sospechosos predefinidos mediante el uso del cálculo de la distancia de similitud. En este trabajo de investigación se incluyó la desambiguación debido a que los recursos se combinaron con la misma entidad produciendo problemas de sinonimia y polisemia.

Para superar el problema de la escasez de datos y la brecha semántica en los textos cortos, se propusieron varios enfoques para agregar semántica al texto contenido en los tweets, uno de ellos fue analizar los hashtags utilizados, los cuales se componen por indicadores sobresalientes que permiten detectar eventos y temas

de tendencias, especialmente para identificar y detectar temas sospechosos y eventos ilegales.

3.1.2 Análisis de delitos: análisis de delitos a través de artículos periodísticos *(Jayaweera & Wijayasiri, 2015).*

El objetivo de este trabajo de investigación fue desarrollar un sistema basado en la web; compuesto por técnicas de análisis de delitos como la detección de puntos de acceso, la comparación de delitos y la visualización de patrones de delitos.

Los datos de análisis que se utilizaron para desarrollar dicho sistema fueron recopilados de periódicos debido a que en ellos se encuentran artículos periodísticos que presentan información sobre incidentes delictivos; dicho sistema se compuso por siete componentes principales los cuales son:

- Rastreadores: encargados de rastrear artículos de noticias en un periódico determinado y almacenarlos.
- Clasificadores: encargado de clasificar los artículos en aquellos que presentan delito y los que no presentan delito.
- Extractores de entidades: encargados de extraer las entidades importantes de los artículos de periódico ya clasificados, tales como fecha del crimen, ubicación, policía, corte y recuento de víctimas.
- Detectores de duplicados: encargados de identificar artículos duplicados y eliminarlos de la base de datos.
- Manejadores de bases de datos: encargados de todas las transacciones de la base de datos.
- Analizadores: encargados de realizar las operaciones de análisis de delitos en los artículos de delitos procesados.
- Interfaces gráficas de usuario: permitieron visualizar los detalles estadísticos del crimen en años anteriores.

La clasificación de este tipo de incidentes se realizó mediante el clasificador de Máquinas de Soporte Vectorial el cual permitió realizar un pre-procesado de los datos para analizar los delitos y los resultados que se obtuvieron de dicho análisis.

Obteniendo como resultado que el sistema propuesto no evalúa los artículos sin antes ser pre-procesados.

3.1.3 ¿Símbolos odiosos o personas odiosas? Características predictivas para la detección del odio en Twitter *(Waseem, 2016)*.

En este trabajo de investigación el objetivo principal fue analizar el impacto de algunas características extralingüísticas en conjunto de n-gramas de caracteres, para la identificación de discursos de odio en comentarios extraídos de la red social Twitter.

Para llevar acabó el análisis de dichas características, se realizó la descarga de 16000 comentarios con los cuales se creó un corpus y se determinó que la longitud de las palabras o áreas geográficas no tienen ningún efecto positivo en el rendimiento al momento de establecer las características extralingüísticas a nivel de carácter.

De igual manera una vez definidos los n-gramas de caracteres de los comentarios se realizó el preprocesado de los mismos, eliminando principalmente Stopwords, Retweets y Signos de puntuación.

Por lo que se prosiguió con la clasificación de los comentarios, obteniendo que el 50.08% de ellos presentaba contenido de racismo y sexismo hacia los hombres y que solo el 2.26% era dirigido hacia las mujeres.

Para la evaluación de la influencia de las características se utilizó un clasificador de Regresión Logística para identificar el rendimiento en cuanto a la utilización de n-gramas de caracteres en donde se obtuvo un 72.87% de Precisión y 77.75% de Cobertura.

3.1.4 Investigando crímenes usando minería de textos y análisis de redes *(Elyezjy & Elhaless, 2015)*.

Este trabajo de investigación se enfocó en desarrollar un sistema para el análisis de documentos basados en investigaciones policiales, con el objetivó principal de identificar posibles delincuentes y las relaciones sociales entre sí a partir de investigaciones de texto no estructurado y extracción de información útil.

El desarrollo de dicho sistema se basó en cuatro actividades base:

- Recopilación de datos
- Pre-procesamiento de datos
- Tokenización
- Normalización

Las cuales permitieron la generación de reglas gramaticales implementadas como expresiones regulares basadas en el conocimiento lingüístico, con el fin de generar patrones para la identificación de ubicación, nombre de la persona y nombre de la organización en documentos de investigación.

Para la evaluación del desempeño y la efectividad del sistema propuesto se consideró el 70% de los documentos para entrenamiento y 30% para pruebas; obteniendo que el sistema es capaz de identificar en textos de investigaciones policiales los nombres de los delincuentes con una precisión del 84% y un 90% en la relación que existe entre ellos.

3.1.5 Análisis de sentimiento en el conjunto de datos de Twitter usando el algoritmo Naïve Bayes *(Parveen & Pandey, 2016)*.

En este trabajo de investigación se llevó a cabo el análisis de publicaciones realizadas por usuarios en Twitter, con el propósito de predecir a través de un marco de Hadoop si dichas publicaciones pertenecen a un sentimiento positivo, negativo o neutral.

El desarrollo de dicho análisis se realizó mediante la extracción publicaciones de Twitter las cuales se pre-procesaron para eliminar: URL´s, caracteres especiales, nombres de usuarios, espacios en blanco, hashtags y realizar la conversión de emoticones.

Después de realizar el pre-procesado de los tweets se aplicó el algoritmo del Naïve Bayes para clasificar las publicaciones en los tres tipos de sentimientos previamente mencionados, una vez clasificadas las publicaciones se hizo uso de un diccionario de palabras compuesto por sinónimos y tipo de polaridad; con el fin de

realizar la implementación de la técnica de Mapeo de frases de acuerdo al tipo de polaridad que presentaba.

Sin embargo, se observó que cuando se hacia la transformación de los comentarios sin considerar emoticones; estos eran ignorados por el algoritmo de Naïve Bayes; por lo que se determinó contemplarlos en el pre-procesado para obtener un mejor resultado de precisión.

3.1.6 Clasificación del texto difamatorio tailandés en las redes sociales *(Arreerard & Senivongse, 2018).*

En este trabajo de investigación se presentó un experimento utilizando dos métodos de aprendizaje automático: Máquinas de Soporte Vectorial y Naïve Bayes; con el objetivo de clasificar textos difamatorios en la lengua tailandesa, basándose en el Lenguaje de características léxicas sintácticas (LSF).

Para el desarrollo de dicha investigación fue necesario crear un diccionario de términos divido en tres grupos: Verbos, Pronombres e Insultos mediante registros de Facebook, los cuales sirvieron para crear datos de entrenamiento extrayendo características para cada uno de los cinco enfoques determinados: palabra n-grama, carácter n-grama, términos específicos, estructura de dependencia y polaridad del sentimiento.

La clasificación de texto se realizó mediante el pre-procesamiento de los datos realizando las tareas de filtrado de abreviaturas y tokenización mediante redes neuronales convolucionales, para extraer características acordes a los enfoques determinados.

Como resultado se obtuvo que las Máquinas de Soporte Vectorial se desempeñan mejor que Naïve Bayes ya que la combinación de los enfoques con palabra n-grama y carácter n-grama proporciona una mejor eficiencia en el manejo de los datos con una precisión de 0.74% mientras que en la polaridad de estos se obtiene un valor de precisión de 0.65% y 0.64% de exactitud.

3.1.7 Contrarrestar la incitación al terrorismo de los perfiles de Twitter en el contexto árabe *(Alghofaili & Almishari, 2018).*

El objetivo de este trabajo de investigación fue desarrollar técnicas basadas en el aprendizaje automático para detectar perfiles de Twitter que inciten al terrorismo.

Para el desarrollo de dicho trabajo se construyó una herramienta que lee y analiza el contenido de las cuentas de Twitter en el idioma árabe de tal manera que detecta con precisión si una cuenta está promoviendo el extremismo religioso y la violencia (terrorismo).

Por lo que se extrajeron datos de un conjunto de 600 perfiles de usuarios, de los cuales 100 incitaban al terrorismo; dichos datos fueron pre-procesados y analizados para determinar las características que los algoritmos de aprendizaje automático utilizaron; los cuales tomaron 70% de los datos para entrenamiento y 30% para pruebas.

Con el desarrollo de dicha herramienta se logró obtener un 87% de exactitud en la identificación de estas cuentas.

3.1.8 Análisis profundo en tweets agresivos mexicanos *(Frenda & Banerjee, 2018).*

En este trabajo de investigación el objetivo principal fue la detección de mensajes que agresivos, mediante el análisis de comentarios extraídos de la red social Twitter.

Por lo que para dicho análisis fue necesario considerar factores lingüísticos como: características lingüísticas, rasgos emotivos de la agresividad y aspectos culturales; dichas características fueron enfocadas a la polaridad de los emoticones utilizados en los comentarios, con sentimiento positivo, negativo y neutral, signos de puntuación y caracterización de mayúsculas.

De igual manera se trabajó por medio de un aprendizaje profundo basado en una red neuronal convolucional, para realizar una comparativa de dos modelos con ingeniería de características y el otro sin él. Por lo que al momento de realizar la evaluación del desempeño del sistema con dichos modelos se utilizó la clase de agresividad obteniendo un 26% de precisión con el modelo basado en

características, por lo que no supero al modelo basado en aprendizaje profundo el cual obtuvo una precisión de 33%.

3.1.9 Optimización de la detección de ciberacoso en Twitter basada en el aprendizaje profundo *(Al-Ajlan & Ykhlef, 2018).*

El objetivo principal de este trabajo de investigación fue detectar el ciberacoso en Twitter enfocándose en el aprendizaje profundo y haciendo uso de un algoritmo de optimización metaheuristica; debido a que dicho algoritmo se basa principalmente en especificar los datos de análisis, eliminando la tarea de extracción y selección de característica y reemplazándolo con vectores de palabras que ayudaron a determinar el conjunto de valores óptimos o casi óptimos para la clasificación.

El desarrollo del enfoque se realizó mediante el uso de una Red Neuronal Convolucional, recolectando un total de 20000 comentarios extraídos de Twitter; después se realizó la limpieza de los datos para eliminar Tweets ruidosos, irrelevantes y duplicados; por último, se usó Glo Ve para observar la similitud de palabras y reconocimiento de entidades nombras.

Una vez realizada dicha actividad de identificación se prosiguió con la implementación del algoritmo de optimización metaheuristica, en donde los resultados obtenidos por dicho algoritmo se evaluaron mediante las medidas de: Exactitud, Precisión y Cobertura.

3.1.10 Análisis de texto para la detección de odio en el habla usando la Red Neuronal de propagación hacia atrás *(Setyadi, Nasrun, & Setianingsih, 2018).*

El objetivo principal de este trabajo de investigación fue clasificar textos extraídos de redes sociales que contienen elementos de discurso de odio o no; mediante el uso de una Red Neuronal de propagación hacia atrás.

Los comentarios fueron extraídos mediante el API de Twitter, la cual es una biblioteca de código abierto que permitió realizar búsquedas por nombres de usuarios, intervalos de tiempo y palabras claves.

Para el desarrollo de la investigación se tomaron de manera aleatoria 1235 comentarios, los cuales fueron etiquetados de manera manual asignándoles una etiqueta con valor de (1) aquellos datos que contenían un discurso de odio y una etiqueta con valor de (0) si los datos no contenían discursos de odio.

Una vez etiquetados los comentarios se realizó el pre-procesado de los mismos, con el objetivo de obtener datos ordenados y estructurados; los cuales fueron contabilizados para identificar el número de palabras que aparecen en cada oración.

Los resultados obtenidos de dicho trabajo de investigación permitieron identificar 652 comentarios con discursos de odio y 583 comentarios sin contenido de odio; para la evaluación del algoritmo utilizado para el desarrollo de la investigación se tomaron el 90% de los datos para entrenamiento y 10% para pruebas, obteniendo un 80.66% de Precisión, 90.07% de Cobertura y 89.47% de Exactitud.

3.1.11 Detección automatizada del discurso de odio hacia la mujer en Twitter *(Şahi, Kılıç, & Sağlam, 2018).*

En este trabajo de investigación el objetivo principal fue construir un modelo de aprendizaje supervisado que identificara y clasificara el odio cibernético hacia la mujer en la red social Twitter; basándose principalmente en la utilización de los algoritmos de clasificación de Máquinas de Soporte Vectorial, Naïve Bayes y Árboles de decisión.

Para el desarrollo de dicho modelo de aprendizaje supervisado fue necesario realizar la extracción de 1288 comentarios, para ser pre-procesados y realizar la creación bigramas, unigramas y trigramas a fin de obtener la frecuencia de los términos.

Una vez desarrollado el modelo se realizó la evaluación del mismo, en donde se obtuvo que el mejor resultado fue proporcionado por las Máquinas de Soporte Vectorial con un 97% de Precisión sobre los otros algoritmos de clasificación utilizados; mostrando que la mayoría de los comentarios clasificados en la clase de odio, presentaban contenido de insultos sexuales y homofóbicos.

3.1.12 Ironía y Sarcasmo: Generación y análisis de un corpus utilizando Crowdsourcing *(Filatova, 2012).*

En este artículo se describe un experimento de generación de corpus mediante la extracción de comentarios generados en la plataforma Amazon, los cuales presentan reseñas con contenido sarcástico sobre sus productos; a dichos comentarios se le realizó un análisis cualitativo y cuantitativo.

El objetivo de este trabajo de investigación se basó principalmente en la detección de ironías y sarcasmos a nivel de oración tomando en cuenta: emoticones y expresiones onomatopéyicas para la risa, puntuación pesada marcas, comillas e interjecciones positivas.

Para aprender patrones de texto sobre los enunciados fue importante hacer una comparación de reseñas sobre un mismo producto a fin de identificar si dichas reseñas presentaban contexto sarcástico o contenido regular (considerado como sin sarcasmo).

El corpus obtenido para el entrenamiento de la identificación semiautomática está conformado por 471 sentencias con contenido sarcástico y 5020 sentencias con contenido no sarcástico; mostrando que una de las características clave del sarcasmo es que muestra palabras positivas, pero expresan opiniones negativas.

Es importante mencionar que el sarcasmo se identifica en dos niveles a nivel de oración y a nivel de documento; aunque se muestra que los términos no son estáticos debido a que sufren variaciones regionales.

3.1.13 Desarrollo de un Servicio Web para determinar la polaridad de textos de redes sociales en español *(Baca, 2014).*

El objetivo de este trabajo de investigación fue describir un método que permite la generación de un corpus basándose principalmente en comentarios de Facebook. Para dicho corpus, se recolectaron comentarios de los cuales se seleccionaron palabras de criterio lingüístico explicito, conocidas como aquellas que contienen un significado emocional (positivo o negativo); estas palabras permitieron delimitar la generación del corpus.

Una vez generado dicho corpus se realizó un análisis de polaridad de los textos, a partir de la implementación de un algoritmo de clasificación automática, utilizando recursos léxicos que permiten la extracción de palabras, frases y emoticones utilizados en Facebook.

Para realizar la clasificación de los comentarios extraídos de Facebook se utilizó una muestra de 400 comentarios los cuales se clasificaron en cinco categorías: positivos, muy positivos, neutrales, negativos y muy negativos.

3.1.14 Una representación ontológica de una taxonomía para el cibercrimen. *(Barn & Barn, 2016).*

En este trabajo de investigación se presenta el desarrollo de una taxonomía para el delito cibernético enfatizando que algunos de los delitos que se cometen de manera tradicional también se ven reflejados en el ciberespacio; debido a la disponibilidad de herramientas tecnologías que permiten a usuarios de internet ejecutar delitos cibernéticos que incluyen estafa, robo, piratería y virus.

Los datos con los que se trabajó en esta investigación están relacionados principalmente en acontecimientos delictivos que ocurrieron en foros tecnológicos.

De igual manera se hace uso del Lenguaje Modelado Unificado para representar ontológicamente la taxonomía mostrando de una manera fundamentada la clasificación del delito cibernético, tomando en cuenta el lenguaje conceptual y las limitaciones que lo acompañan para describirlos.

3.1.15 Taxonomía de la comunicación violenta y el discurso del odio en Internet *(Llinares, 2016)*.

En este trabajo de investigación el objetivo principal fue llevar a cabo una observación y análisis de más de 250,000 tweets publicados en el idioma español; dichos comentarios tienen que ver con acontecimientos de ataques ocurridos en París; con el fin de describir las diferentes formas de comunicación violenta existentes en Internet sobre las que gira la discusión social acerca de si deben ser criminalizadas o no, y conceptualizar las mismas en una taxonomía básica que permita identificar cada una de las categorías y clasifique las diferentes formas de comunicación violenta entre las cuales está el discurso del odio.

Para esto fue necesario realizar las siguientes actividades:

a) Delimitar el objetivo de la observación y la categorización de la comunicación violenta y el discurso de odio.
b) Observar los mensajes a partir de una muestra representativa.
c) Construir la taxonomía basándose en lo observado y apoyada en los criterios valorativos de interés; los cuales se enfocan principalmente en daños físicos y daños morales (ofensa).

Una vez realizadas dichas actividades se realizó un análisis de una muestra de 200 tweets utilizando a un grupo de interjueces para comprobar que había una alta fiabilidad en la valoración de todos y cada uno de ellos; esto se llevó a cabo mediante el uso del índice de concordancia de Kappa; el cual menciona que cuando su resultado arroja un valor de k inferior a 0.20 la fuerza de concordancia es pobre, entre 0.21 y 0.40 es débil; de 0.41 a 0.60 moderada, de 0.61 a 0.80 buena y entre 0.81 y 1 muy buena.

Como resultado de dicho análisis de la muestra se obtuvo que el nivel de concordancia de interjueces fue de 0.91 por lo que se consideró más que satisfactorio.

3.1.16 Análisis de Sentimientos para la prevención de mensajes de odio en las Redes Sociales *(Montoro, 2019).*

En este trabajo de tesis se utilizó el análisis de sentimiento centrándose principalmente en la polaridad de las expresiones plasmadas en texto, etiquetando su conjunto de datos como positivo o negativo con diferentes grados de pertenencia a fin de determinar si un mensaje de odio se desprende un sentimiento negativo.

De igual manera se describe la creación de una taxonomía compuesta principalmente por discursos que presentan una comunicación violenta y de odio. El objetivo de la taxonomía fue identificar mensajes de odio que inciten a realizar actividades de violencia o injurias contra la sociedad.

Para el desarrollo de dicha taxonomía fue necesario establecer una búsqueda de patrones que permitieran identificar de manera precisa mensajes de Comunicación

violenta y Discurso de odio; tomando en cuenta los parámetros que componen un Delito de Odio; con el objetivo de establecer en un futuro un mecanismo computacional que sea capaz de identificar dichos mensajes.

Dicha taxonomía permite distinguir dos clases bien diferenciadas, la identificación del Discurso de Odio propiamente dicho (con la detección de incitación a la violencia e injurias) y los agravantes, divididos a su vez, en agravantes propios del mensaje, del entorno y del clima.

3.1.17 Tabla de trabajos representativos

A continuación, en la Tabla 3 se presentan algunos de los trabajos más representativos encontrados en la literatura, los cuales fueron seleccionados con el propósito de identificar las actividades desarrolladas en el análisis de violencia en texto.

Tabla 3: Trabajos Representativos sobre el Análisis de violencia en texto

Artículo	Año	Objetivo	Datos Analizados	Método/Algoritmo utilizado	Métrica de evaluación
¿Símbolos odiosos o personas odiosas? Características predictivas para la detección del odio en Twitter.	2014	Analizar el impacto de características extralingüísticas en conjunto de n-gramas de caracteres, para la identificación de discursos de odio en comentarios extraídos de la red social Twitter.	Comentarios extraídos de la red social Twitter	Clasificador de Regresión Logística	72.87% de Precisión 77.75% de Cobertura
Análisis de sentimiento en el conjunto de datos de Twitter usando el algoritmo Naïve Bayes.	2016	Identificar mediante un marco de Hadoop si publicaciones generadas en Twitter pertenecen a un sentimiento positivo, negativo o neutral.	Publicaciones de la red social Twitter	Naïve Bayes	71.12 % de Precisión
Clasificación del texto difamatorio tailandés en las redes sociales.	2018	Clasificar textos difamatorios en la lengua tailandesa, basándose en el Lenguaje de	Textos difamatorios	Máquinas de Soporte Vectorial	74% de Precisión

		características léxicas sintácticas (LSF).		Naïve Bayes	
Contrarrestar la incitación al terrorismo de los perfiles de Twitter en el contexto árabe.	2018	Desarrollar técnicas basadas en el aprendizaje automático para detectar perfiles de Twitter que inciten al terrorismo.	Información sobre 600 perfiles de usuarios de la red social Twitter	Algoritmo de aprendizaje automático	87% de Exactitud
Análisis de texto para la detección de odio en el habla usando la Red Neuronal de propagación hacia atrás	2018	Clasificar textos extraídos de redes sociales que contienen elementos de discurso de odio o no; mediante el uso de una Red Neuronal Convolucional.	1235 comentarios extraídos de la red social Twitter	Red Neuronal Convolucional	66% de Precisión 90.07% de Cobertura 89.47% de Exactitud.

3.2 Aplicaciones de Lógica Difusa en análisis de texto

Dentro del estudio de la literatura se identificaron algunos artículos que presentaban Algoritmos Difusos desarrollados para identificar patrones que se presentan en redes sociales, donde dichos patrones permiten determinar la conducta que los usuarios presenta por medio de la información que comparten; los autores *(Romsaiyud & Premchaiswadi, 2015)* lograron dicha identificación mediante la extracción de reglas semánticas al igual que los autores *(Mukkamala, Abid, & Vatrapu, 2014).*

3.2.1 Análisis de sentimiento basado en conjuntos difusos de grandes datos sociales *(Mukkamala, Abid, & Vatrapu, 2014).*

En este trabajo de investigación se presentó la discusión de un modelo conceptual de datos sociales; mediante el cual se realizó un análisis de un modelo formal basado en conjuntos Fuzzy, describiendo reglas semánticas que se obtuvieron de comentarios publicados por usuarios de la red social Facebook.

Para el desarrollo de dicho modelo fue necesario hacer uso de la herramienta SODATO, la cual permitió obtener datos de Facebook referente a un tema en específico y realizar la clasificación sentimental que presentan dichas publicaciones.

Es por ello que mediante el análisis de sentimientos basado en la teoría de conjuntos difusos se utilizó un marco de grandes datos sociales, para entender al usuario tan solo con revisar el contenido de sus publicaciones y así determinar cuándo una publicación presenta contenido positivo, negativo o neutral.

3.2.2 Aplicación de técnicas de minería de patrones secuenciales difusos para predecir el liderazgo en las redes sociales *(Romsaiyud & Premchaiswadi, 2015).*

En este artículo se presentó un algoritmo nuevo de minería difusa, el cual permitió encontrar patrones difusos secuenciales de los miembros de distintos grupos redes sociales; integrando conceptos de conjuntos difusos, por lo que para su desarrollo se hizo uso del Análisis de redes sociales (SNA) para mapear y medir las relaciones

entre los distintos individuos pertenecientes a un grupo; permitiendo obtener una secuencia de patrones de comportamiento.

De igual manera se realizó un análisis de texto basado en un conjunto de teorías y tecnologías para realizar la representación sintáctica de los textos.

El aporte de este artículo fue la creación de un marco que mediante la extracción de reglas predice el comportamiento de los usuarios de un grupo en específico, mediante el análisis de sus publicaciones generadas en la red social Facebook; identificando en ellas la presencia de emociones positivas o negativas con el fin de agruparlas y sumarlas dependiendo al tipo de emoción que presenten.

3.2.3 Clasificación de los comentarios de las redes sociales mediante una técnica de clasificador de base difusa *(Bairagi & Tapaswi, 2016).*

En este trabajo de investigación se enfocaron en estudiar cuales son las técnicas que permiten identificar patrones generados por usuarios y a su vez como estos influyen en otros usuarios; mediante una clasificación difusa permitiendo obtener resultados más precisos.

Es importante mencionar que el objetivo principal de este trabajo fue desarrollar un clasificador de texto binario basado en reglas difusas para analizar textos extraídos de redes sociales e identificar en ellos la presencia de engaño social o bulling.

Dicho clasificador fue desarrollado en dos módulos de datos, el primero se enfocó en aprender de patrones de comunicación anteriores o históricos disponibles y el segundo solo fue creado para pruebas; de igual manera se realizaron actividades como: eliminación de Stopwords, análisis de datos, normalización de datos, desarrollo de reglas difusas y creación de mensajes de decisión.

3.2.4 Detección y prevención de delitos mediante análisis de redes sociales *(Gupta & Kumar, 2015)*

En este trabajo de investigación el objetivo principal fue analizar el comportamiento de los datos obtenidos de un conjunto de actividades involucradas

en delitos, mediante técnicas de redes neuronales que permitan implementar el proceso de sincronización de la información.

En cuanto al desarrollo de dicha investigación se utilizó el Modelo Kuramoto el cuál es un modelo matemático que se enfoca en describir la sincronización de las redes mediante la observación de las variaciones que estas presentan al interactuar entre sí; de igual manera se utiliza el Modelo de Barabási-Albert para analizar la vinculación de cada una de las redes; también fue necesario hacer uso de técnicas de un sistema difuso para realizar la transformación de los datos de entrada al dominio previamente definido, el cual se compone por las reglas difusas de forma canónica.

Como resultado de dicha investigación se obtuvo la clasificación de los niveles de criminalidad que presentan al momento de interactuar entre sí, basándose en reglas previamente definidas las cuales permitieron realizar la clasificación de los mismos.

La evaluación del modelo de predicción generado se llevó a cabo mediante al análisis de 3 algoritmos: Random Forest, Regresión Lineal y Máquinas de Soporte Vectorial, utilizando la métrica de Precisión.

El clasificador desarrollado fue evaluado contra los clasificadores de Bayes y KNN por medio de la métrica de exactitud, obteniendo como resultado que el clasificador basado en reglas difusas obtuvo un porcentaje de 78.32 mejor que los otros clasificadores.

3.2.5 Enfoque de modelado de temas difusos para la minería de texto sobre texto breve (Rashid &Shah, 2019)

El objetivo de este trabajo de investigación está enfocado en presentar un nuevo enfoque de modelado de temas difuso para el análisis de textos breves para mejorar el ruido y la escasez en los textos cortos; identificando en ellos términos locales o globales mediante el cálculo de la frecuencia de aparición de términos; la cual se calcula a través de un modelo de bolsas de palabras en donde cada una de ellas presentan un grado de pertenencia.

En el desarrollo de este trabajo se utilizó el algoritmo Fuzzy c-means, el cual es el encargado de eliminar el impacto negativo en la ponderación global de términos permitiendo así la semántica de los temas relevantes que conforman dichos textos.

Los resultados obtenidos con dicho modelo fueron favorables en cuanto a la identificación de los términos que conforman a los textos breves, obteniendo así una precisión del 0.82.

3.2.6 Trabajos representativos

A continuación, en la Tabla 4 se presentan algunos de los trabajos más representativos encontrados en la literatura, los cuales muestran la creación de reglas difusas que conforman a los modelos basados en lógica difusa, con el propósito de identificar cuáles son los pasos y características a considerarse para su creación.

Tabla 4: Trabajos Representativos sobre el uso de Lógica Difusa

Artículo	Año	Objetivo	Datos Analizados	Utilidad
Análisis de sentimiento basado en conjuntos difusos de grandes datos sociales	2014	Describir un modelo formal basado en la teoría de conjuntos difusos para analizar datos de la red social Facebook y clasificarlos mediante el análisis de sentimientos.	Publicaciones y comentarios generados en la red social Facebook.	Información sobre el uso de conjuntos difusos para la clasificación de datos.
Aplicación de técnicas de minería de patrones secuenciales difusos para predecir el liderazgo en las redes sociales	2015	Presentar un algoritmo nuevo de minería difusa utilizando técnicas que permitan el análisis de redes sociales para identificar patrones difusos.	Publicaciones generados en la red social Facebook.	Información sobre la creación de reglas difusas que permiten clasificar publicaciones dependiendo de una emoción positiva o negativa.

Clasificación de los comentarios de las redes sociales mediante una técnica de clasificador de base difusa.	2016	Desarrollar un clasificador de texto binario basado en reglas difusas para analizar textos extraídos de redes sociales e identificar en ellos la presencia de engaño social o bulling.	Textos extraídos de redes sociales.	Información sobre la creación de reglas difusas y el desarrollo de un clasificador binario.
Detección y prevención de delitos mediante análisis de redes sociales.	2015	Analizar el comportamiento de los datos obtenidos de un conjunto de actividades involucradas en delitos, mediante el uso de técnicas difusas las cuales permiten transformar los datos de entrada al dominio definido.	Textos sobre actividades involucradas en delitos.	Información sobre el uso de reglas difusas que permiten llevar a cabo la clasificación de datos.
Enfoque de modelado de temas difusos para la minería de texto sobre texto breve.	2019	Presentar un modelo difuso para el análisis de textos breves, identificando en ellos términos locales o globales mediante el cálculo de la frecuencia de aparición	Textos cortos extraídos de redes sociales. Conjunto de preguntas generadas en redes sociales.	Información sobre el uso de la lógica difusa para identificar términos relevantes en textos.

Capítulo 4 Método para la detección de apología del delito

En este capítulo se describe el desarrollo de un modelo para la detección de apología del delito en redes sociales, mediante un sistema difuso basado en el modelo de Takagi-Sugeno.

4.1 Descripción general del método para la detección de apología del delito

En la Figura 4 ,se observa cuál es el método desarrollado para dar solución al tema de investigación “Detección de apología del delito en redes sociales utilizando un modelo difuso”.

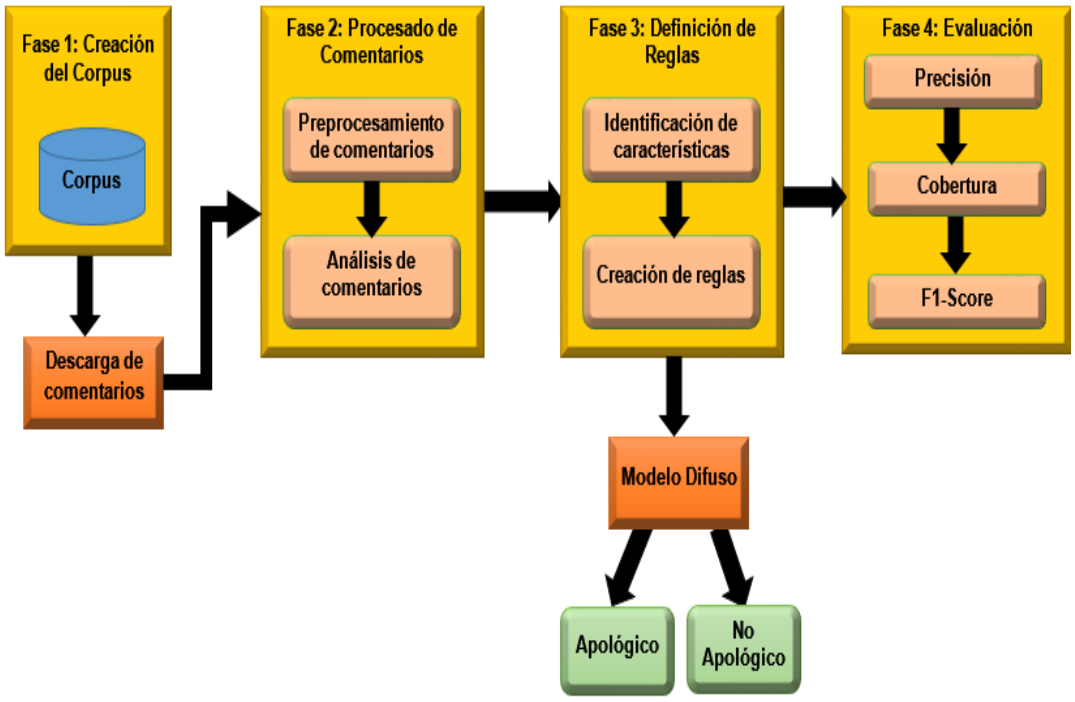


Figura 4: Método para la detección de apología del delito

A continuación, se describen las fases que conforman al método de solución mostrado anteriormente.

Fase 1: Creación del Corpus

El objetivo principal de esta fase fue generar un Corpus Apológico mediante la extracción de comentarios de la red social Twitter, dicho corpus se compone por comentarios que presentan contenido apológico de incitaciones al odio e incitaciones a la violencia.

Fase 2: Preparación de comentarios

En esta fase el objetivo principal fue preparar los datos descargados del corpus enfocado en redes sociales, mediante el preprocesado de los mismos para realizar un análisis en donde se identifiquen cuáles son características obtenidas de los elementos que componen un comentario.

Fase 3: Definición de reglas

En esta fase el objetivo principal fue seleccionar las variables de entrada y salida que permitirán crear el conjunto de reglas difusas que forman parte del modelo difuso a desarrollarse; dichas reglas son fundamentales para poder identificar si un comentario presenta o no contenido apológico.

Fase 4: Evaluación

En esta fase el objetivo principal fue llevar a cabo la evaluación de las fases anteriores del método propuesto de solución mediante las métricas de evaluación: Precisión, Cobertura y F1-Store.

4.2 Creación de un corpus apológico

Esta fase se llevó a cabo debido a que, en investigaciones realizadas en la literatura, no se encontraron conjuntos de textos relacionados con apologías del delito; dicho corpus fue utilizado como base de conocimiento, ya que está compuesto principalmente por comentarios apológicos generados por usuarios de la red social Twitter, los cuales fueron extraídos para ser analizados y determinar si

presentan algún tipo de información relevante en cuanto a la presencia de incitaciones de odio y violencia.

Se realizó un estudio sobre los corpus dentro de la literatura que permiten identificar cuál es el método a seguir para el desarrollo de un corpus y a su vez determinar qué tipo de corpus se utilizará dependiendo de las necesidades del tema de investigación.

Hoy en día se considera que los corpus deben cumplir los siguientes requisitos:

1. *Formato electrónico*

Un corpus, para ser una herramienta útil al lingüista, debe estar informatizado, es decir, los textos de que consta tienen que estar en formato electrónico.

2. *Autenticidad de los datos*

Los textos recogidos en el corpus deben de ser muestras reales de uso de la lengua objeto de estudio.

3. *Criterios de selección*

Los textos que forman parte del corpus deben haber sido elegidos de acuerdo a determinados criterios lingüísticos y/o extralingüísticos.

4. *Representatividad*

Debe de responder a parámetros estadísticos que garanticen que los textos 'representan' la variedad de lengua objeto de estudio ('muestra representativa').

5. *Tamaño*

Los corpus constan de un tamaño finito, que se suele medir en millones de palabras o de formas y que se fija antes de empezar la recogida de los textos; una vez alcanzado ese número, se da por terminada la recopilación del corpus, que no es más que el primer paso de todo el proceso.

Para la generación del corpus fue necesario determinar los tipos de apología que deben presentar los comentarios que conformarán al corpus; en este trabajo de investigación nos enfocamos principalmente en dos clases apológicas:

a) Incitación al Odio

En la cual se pueden encontrar conductas que solo se mantienen en un sentimiento afectando psicológicamente a los individuos, dichas conductas son:

- Amenazas
- Acoso
- Machismo

Contenido que involucre: raza, origen étnico, nacionalidad, orientación sexual, género, identidad de género, afiliación religiosa y discapacidad u enfermedad.

b) Incitación a la violencia

En las cuales podemos encontrar conductas como:

- Deseos de muerte
- Daños físicos
- Enfermedades
- Muerte
- Actos de terrorismo

Las cuales, a diferencia de la incitación al odio, estas involucran una acción que puede afectar físicamente a los individuos.

A continuación, se describen a detalle el desarrollo del método para la generación del corpus con comentarios extraídos de la red social Twitter.

La construcción del corpus se realizó mediante el método de *Aguado de Cea (2012)*, el cual está conformado por las etapas de *Recopilación de textos de salida, Determinación de las entradas, Análisis de los textos recopilados y de las entradas, Construcción de un corpus, y Extracción de patrones*, tal como se observa en la Figura 5. Sin embargo, para la generación del corpus apológico solo se consideraron las cuatro primeras; la etapa de *Extracción de patrones* solo se considera si se pretende reutilizar el trabajo.

Figura 5: Método para generar un corpus (Aguado de Cea, 2012)

4.2.1 Recopilación de textos

La primera etapa del método es la recopilación de textos, la cual tiene como objetivo adquirir una gran cantidad de información. No importa tanto su validez, sino que den una muestra completa de todas las posibilidades que se pueden presentar como texto de salida aceptables.

La fuente de información seleccionada para la generación del corpus apológico fue la red social Twitter, de donde se extrajeron comentarios generados por los usuarios de esta red social, considerados como textos de salida.

RStudio fue la herramienta que se utilizó para la extracción de comentarios mediante una aplicación generada en la Plataforma de Desarrolladores de Twitter tal como se observa en la Figura 6 través de una cuenta personal de usuario; esto con el propósito que por medio de la aplicación se obtuvieran los permisos para trabajar con los datos de los usuarios pertenecientes a esa red social y con el contenido que estos comparten.

Figura 6: Aplicación generada en la Plataforma de Desarrolladores de Twitter (Twitter Developer, 2019)

La creación de la aplicación en la plataforma de desarrolladores, genera automáticamente dos claves únicas de conexión: Claves de la API de consumidor y Acceso a tokens secretos como se muestran en la Figura 7, las cuales permiten la extracción de información de la plataforma de Twitter.

Figura 7: Claves y acceso a tokens(Twitter Developer, 2019)

Una vez obtenidas esas claves se prosiguió a seleccionar la herramienta de RStudio debido a que mediante el paquete de twitteR se proporciona acceso a la API de Twitter para extraer los comentarios.

Las funciones que se pueden realizar con twitteR son las siguientes:

a) Funciones para trabajar con retweets
b) Funciones para ver las tendencias de Twitter
c) Funciones para descargar comentarios por fecha, lenguaje o termino.

La Figura 8 muestra la realización del script para la descarga de los comentarios y su exportación a un documento tipo csv; en dicho script se determinan los criterios que se consideran en a descarga; los cuales son los siguientes:

a) Latitud
b) Altitud
c) Radio cobertura en Kilómetros (Km). Lo cual representa el lugar de procedencia de donde se realizará la descarga
d) Número de comentarios a descargar

e) Periodo de fecha

f) Lenguaje en que se encuentran escritos los comentarios.

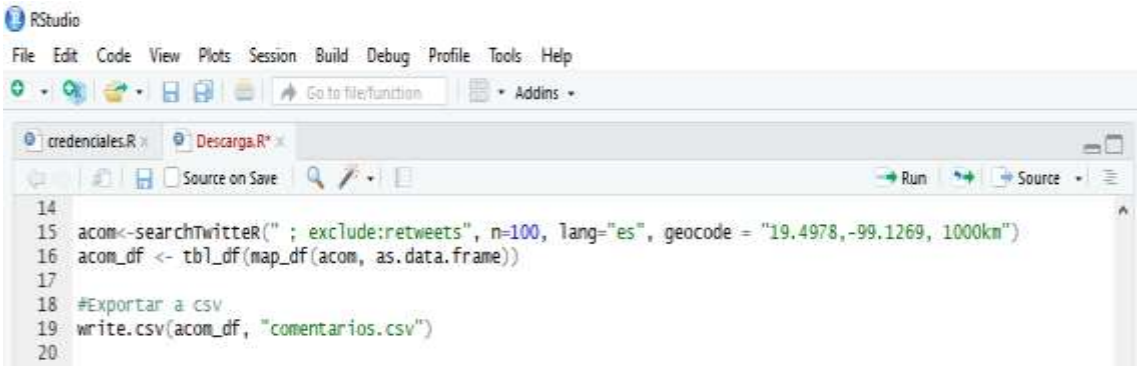


Figura 8:Script en RStudio

De igual manera la extracción de comentarios se realizó mediante un proceso de recolección como se muestra en la Figura 9 en dicho proceso se realizaron varias descargas, algunas de ellas incluyendo palabras clave determinadas previamente por el análisis previo de un número de comentarios considerables en el cual se determinó la frecuencia con la que aparecían dichas palabras. En el **Anexo A** se muestra un listado de las palabras claves utilizadas para la búsqueda y extracción de comentarios.

Figura 9: Proceso de recolección de comentarios

4.2.2 Determinación de posibles entradas

En esta etapa, se seleccionaron los posibles textos de entrada aceptados por el sistema. Las entradas deben de estar bien definidas para evitar contenido que no aporte nada a la generación del corpus.

Los textos de entrada son considerados como aquellos comentarios que fueron seleccionados porque presentan contenido apológico en su contexto, y contenían información referente a clases de incitación al odio e incitación a la violencia.

La estructura del corpus es: Número de comentario (No.), Texto del comentario (Comentario), Fecha de realización del comentario (Fecha), Clase apológica (Clase); tal como se observa en la Tabla 5.

Tabla 5: Estructura del Corpus

No.	Comentario	Fecha	Clase
1	@NicolasMaduro Muérete maldita piltrafa!!! Eres la peor mierda que ha pasado por este mundo!!	19/03/2019	Odio
2	@RoxSanchez10 No te preocupes tanto por AMLO, tú también llegarás a la muerte y qué será de ti…?	19/03/2019	Odio
3	¿Si eres ateo qué te impide salir a matar y violar todo lo quieras?	19/03/2019	Violencia
4	Ojala los encuentren y los quemen vivos a esos desgraciados ladrones mal paridos cárcel no merecen	19/03/2019	Odio
5	Malditos corruptos, matenlos!!!	19/03/2019	Violencia

El corpus está compuesto por 4,600 comentarios extraídos de la red social Twitter en diferentes intervalos de fecha que se encuentran dentro del periodo 01/03/2019 al 30/08/2019.

4.2.3 Análisis de los textos y de las entradas

El objetivo de esta etapa fue conseguir una comprensión detallada de las correspondencias entre los textos de salida y los datos de la fuente de información.

El estudio de las relaciones entre los textos de salida y las entradas, servirá para establecer las correspondencias entre ambos al igual que para descubrir textos sin ninguna entrada asociada y viceversa.

Asimismo, en esta etapa se realizó el preprocesado de los comentarios con el fin de obtener una limpieza de los mismos, por lo que para esto fue necesario utilizar el corpus ya generado, solo los datos correspondientes a la columna de Comentarios.

4.2.3.1 Preprocesado de comentarios del Corpus

El preprocesamiento de datos engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, entre otras.) Con el preprocesamiento de datos se pretende que los datos a ser utilizados en tareas de análisis o descubrimiento de conocimiento conserven su coherencia *(Zhang, Zhang, & Yang, 2003).*

En este trabajo de investigación se recibió como entrada un texto sin formato que contiene comentarios que componen el Corpus Apológico; en la Figura 10 se pueden observar las técnicas utilizadas para llevar a cabo el preprocesamiento de los comentarios.

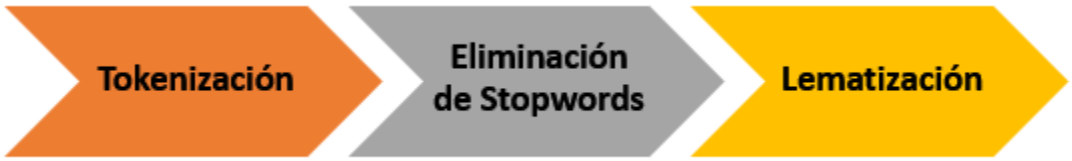


Figura 10: Técnicas utilizadas en el preprocesamiento de comentarios

A continuación, se describen brevemente cada una de estas técnicas aplicadas a los comentarios extraídos:

- **Tokenización**

(Trim, 2018) describe el proceso de segmentar el texto en palabras y oraciones llamados *tokens*.

En la implementación de esta técnica, se trabajó con la columna de comentarios, con el objetivo de que dichos comentarios extraídos de la red social twitter pasaran por el proceso de tokenización obteniendo alrededor de 36, 000 tokens.

- **Eliminación de Stopwords**

(Landaeta, 2014) menciona que la eliminación de Stopwords consiste en eliminar las palabras que no aportan significado o información relevante, como lo son: artículos, pronombres, algunos verbos, conjunciones, etc.

En la implementación de esta técnica en el preprocesado de los comentarios, se cargó automáticamente una lista de Stopwords, en la cual a partir de las palabras obtenidas en el proceso de tokenización se identifican y eliminan cada una de las palabras que se encuentran en la lista.

(Montesinos, 2014) proporciona una lista de Stopwords utilizada para el preprocesado de los comentarios la cual puede verse en el **Anexo B**; dicha lista fue analizada y previamente modificada tomando en cuenta aquellas palabras que pueden ser consideras en el análisis de los comentarios las cuales se muestran en la Tabla 6.

Es decir que si se eliminan estas palabras, el contexto de los comentarios cambiaria tal como se muestra en la Tabla 7 y, por lo tanto, no se lograría identificar con certeza si un comentario está presentando o no contenido apológico.

Tabla 6: Stopwords no eliminadas en el análisis de comentarios

Adverbio, preposiciones y verbos	
si	son
contra	usamos
entre	vamos
eres	tendrás
hacemos	están

Tabla 7: Ejemplo

Texto original	Vamos a matar a los homosexuales y así hacemos un bien a la sociedad
Texto con preposiciones y Stopwords	Vamos matar homosexuales hacemos bien sociedad
Texto sin preposiciones ni Stopwords	matar homosexuales bien sociedad

- **Lematización**

La lematización consiste en determinar la forma de una palabra que se constituye en *lema (Sánchez, 2017).* Otros investigadores como *(Gómez, 2005),* la definen como el proceso de eliminación automática de partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (*lema*).

Se utilizó la librería *Freeling* para lematizar los comentarios con el objetivo de obtener la forma base de las palabras que componen a dichos comentarios. En la Tabla 8 se observa un ejemplo de los comentarios ya lematizados.

Tabla 8: Comentarios lematizados

No.	Comentario
1	nicolasmaduro muérete maldito piltrafa eres peor mierda pasado mundo
2	roxsanchez preocupar amlo también llegar muerte ser buen prianista corrupto ir directamente infierno
3	si ser ateo impedir salir matar violar querer
4	ojalar encontrar quemar vivo desgraciado ladrón mal parir cárcel no merecer ser quemar
5	maldito corrupto matar

4.2.4 Construcción de un corpus

A partir del análisis efectuado en la etapa anterior, y de los criterios previamente establecidos se proseguirá con la generación del corpus, el cual debe de ser correcto (gramaticalmente y dentro del dominio) pero a la vez contar con la capacidad de ser modificado si en algún momento es necesario eliminar o agregar algún texto.

Sin embargo antes de continuar con la construcción del corpus fue necesario realizar a creación del Perfil de un analista de contenido el cual se creó con el propósito de identificar el perfil idóneo del individuo que realizará la actividad de crear un corpus, mediante el análisis de los datos extraídos de la red social Twitter por lo que fue necesario identificar los Conocimientos, Funciones y Habilidades que dicho individuo debe de poseer y cubrir tal como se observa en la Tabla 9.

Un analista de contenido web, es el encargado de medir y analizar los efectos de cualquier acción realizada en Internet; su función principal es recopilar y analizar datos que se encuentran en la web con el fin de interpretar dichos datos estableciendo relaciones de causa-efecto respecto a las acciones tomadas *(Mañes, 2018).*

Tabla 9: Aspectos de un Analista de Contenido (Autoría propia, 2019)

Conocimientos	Funciones	Habilidades
• Uso de herramientas informáticas. • Uso de API´s para la extracción de contenido en Twitter. • Dominio del lenguaje de programación Python para el procesamiento de comentarios en Twitter. • Conocimiento del lenguaje coloquial empleado en México.	• Búsqueda de información en redes sociales. • Monitoreo de medios digitales. • Filtrar y seleccionar los datos que representen amenazas. • Gestión de redes sociales. • Análisis de datos. • Discernir entre publicaciones valiosas e irrelevantes.	• Sólidas habilidades de investigación y una actitud excepcional sobre el trabajo en equipo. • Disponibilidad de horario. • Dedicación. • Adaptabilidad. • Responsabilidad. • Identificar ambigüedad en comentarios. • Capacidad de razonamiento y habilidades analíticas en diversas situaciones.

Para poder realizar el análisis de la información que se extrajo de la red social Twitter se utilizaron cinco individuos, los cuales ejercen la carrera de Ingeniería en Sistemas Computacionales y cuentan con los conocimientos necesarios en el uso de herramientas informáticas y procesamiento de datos.

Para que dichos individuos fueran aceptados como personas idóneas para realizar dicha tarea, fue necesario realizar una pequeña capacitación en cuanto al contenido que iban a identificar en la información extraída de la red social Twitter; en donde dicha información está basada sobre acontecimientos que comúnmente se ocurren hoy en día y que pueden ser identificados mediante experiencias propias de los individuos.

De igual manera es importante mencionar que cualquier individuo relacionado con la el área de computación cumple con el perfil solicitado para realizar actividades relacionadas con el análisis de la información.

Una vez definido dicho perfil se continuo con la creación del corpus, de tal manera que se recopilaron todos los comentarios que fueron preprocesados de manera correcta en un archivo de texto sin formato (Tabla 10) los cuales se utilizaron en la anotación del corpus.

Es importante enfatizar que el corpus debe ser representativo con respecto al dominio de delitos apológicos y contar con la capacidad de ser modificado si en algún momento es necesario eliminar o agregar algún texto.

Tabla 10: Corpus apológico preprocesado

No.	Comentario	Fecha
1	nicolasmaduro muérete maldito piltrafa eres peor mierda pasado mundo	19/03/2019
2	roxsanchez preocupar amlo también llegar muerte ser buen prianista corrupto ir directamente infierno	19/03/2019
3	si ser ateo impedir salir matar violar querer	19/03/2019
4	ojalar encontrar quemar vivo desgraciado ladrón mal parir cárcel no merecer ser quemar	19/03/2019
5	maldito corrupto matar	19/03/2019

Es importante mencionar que al momento de construir el Corpus se identificaron ciertas palabras clave que servirán para futuras extracciones de comentarios, están se observan en el **Anexo A.**

4.2.4.1 Validación del corpus

Una vez generado el corpus de comentarios apológicos, se implementó el proceso de validación del mismo a fin de poder determinar si dicho corpus cumple con las características necesarias para ser utilizado, es decir, si cuenta con comentarios que contengan alguna de las palabras que pertenecen al conjunto de palaras clave identificadas en el análisis de comentarios con contenido apológico y

a su vez sustentada con el corpus Léxico de intensidad de emoción *(Mohammad, 2016)*, el cual presenta términos prominentes en las redes sociales.

Para esto fue necesario primero realizar un Proceso de Capacitación tal como se muestra en la Figura 11, el cual se impartió a los etiquetadores que cumplieron con dicha actividad de etiquetado.

Figura 11: Proceso de capacitación

Los comentarios fueron proporcionados en un archivo tipo csv, a un grupo de cinco individuos que fungieron como etiquetadores.

Se consideraron dos clases para el etiquetado:

- Apológico (Comentarios con contenido apológico)
- No apológico (Comentarios sin contenido apológico)

Por lo que los etiquetadores llevaron a cabo la lectura de cada uno de uno de los comentarios y de acuerdo a su criterio personal determinaban a que clase pertenecían dichos comentarios.

A continuación, se describen las seis etapas que conforman el Proceso de capacitación; las cuales son: Definición de Apología del delito, Tipos de clases, Palabras Clave, Tipos de incitaciones, Tipos de contenido y Ejercicios de prueba.

1. Definición de Apología del delito

A los individuos que realizaran la tarea del etiquetado de comentario se le presentara primero la definición de lo que es Apología del delito; con el propósito de que ellos identifiquen las conductas que se ven involucradas con dicho concepto.

La definición que se les presento fue la siguiente:

Apología proviene del latín apologia y significa “discurso en defensa o alabanza de persona o cosa” y delito proviene del latín delicto o significa “culpa, crimen o

quebrantamiento de la ley" por lo que el significado en su conjunto es el de: alabanza de un quebrantamiento grave de la ley ***(****Diccionario Jurídico Mexicano).*

2. Tipos de clases

Se les presentaron tres tipos de clases a identificarse que se deben de considerar al momento de leer el contenido de los comentarios, para asignarle una de las siguientes clases:

- **Apológico:** Comentarios que los etiquetadores consideran a criterio propio que presentan contenido apológico.
- **No apológico:** Comentarios que los etiquetadores consideran a criterio propio que no presentan contenido apológico.
- **Indeciso:** Comentarios en que los etiquetadores no están del todo seguros si presentan o no contenido apológico.

3. Palabras Clave

A los etiquetadores se les entregará un listado de las palabras clave consideradas en comentarios que presentan contenido apológico. Estas palabras fueron obtenidas de un análisis previo realizado en comentarios descargados de la red social Twitter y a su vez sustentada con el corpus Léxico de intensidad de emoción (Mohammad, 2016), el cual presenta términos prominentes en las redes sociales.

Las palabras clave las podemos observar en el **Anexo A** de este trabajo de investigación.

Es importante mencionar a los etiquetadores que deben de identificar las palabras clave en el contenido de los comentarios que van a etiquetar; esto con el fin de definir si el comentario está presentando un contenido apológico. Sin embargo, también deben de considerar que no por el simple hecho de que un comentario presente una palabra que se encuentre dentro de la lista de palabras clave quiere decir que es un comentario apológico; esto debe de definirse dependiendo del contexto que este presentando el comentario.

4. Tipos de Incitaciones

En cuanto al tipo de incitaciones que pueden encontrarse dentro de un comentario que presente contenido apológico se seleccionaron principalmente dos tipos, las cuales son las siguientes:

- **Incitaciones de Odio**

Son aquellas conductas que solo se mantienen en un sentimiento afectando psicológicamente a los individuos, como lo son:

  - Amenazas
  - Acoso
  - Machismo

Así como contenido que involucre: raza, origen étnico, nacionalidad, orientación sexual, género, identidad de género, afiliación religiosa y discapacidad u enfermedad.

- **Incitaciones de Violencia**

Se pueden encontrar conductas como:

  - Deseos de muerte
  - Daños físicos
  - Enfermedades
  - Muerte
  - Actos de terrorismo

Las cuales, a diferencia de la incitación al odio, estas involucran una acción que puede afectar físicamente a los individuos.

5. Tipos de Contenido

En esta etapa se les presentará a los etiquetadores cuatro tipos de contenido que pueden tener los comentarios. Estos contenidos se obtuvieron del artículo "Taxonomía de la comunicación violenta y el discurso del odio en Internet" *(LIinares, 2016).*

Dicho artículo es considerado debido que se enfoca en llevar a cabo el etiquetado de comentarios que presenten contenido de discurso de odio o comunicación violenta.

Los contenidos son los siguientes:

1. La voluntad de realización directa, o a través de otros a los que se incite directamente, de actos de violencia física contra personas concretas o indeterminadas, así como expresiones de referencia en positivo (en forma de defensa, enaltecimiento, justificación, banalización, comprensión, alegría) a la acusación de tal violencia.

2. El insulto o la ofensa grave dirigida a personas concretas y determinadas, así como la atribución a estas de la realización de hechos delictivos o ilícitos graves con conocimiento de la falsedad o con temerario desprecio hacia la verdad.

3. El desprecio o expresión de odio hacia grupos determinados, especialmente hacia aquellos que de algún modo han visto, o pueden ver, privados sus derechos y que sufren actividades intolerantes, y en particular aquellas expresiones que usen términos despectivos contra los mismos y que pidan o justifiquen la restricción de derechos contra tales grupos.

4. Aquellas expresiones especialmente desagradables y de muy mal gusto referidas a sucesos que causan grave dolor a algunas personas, en particular las que muestran odio a dichas personas o las que deshumanizan totalmente al que las realiza, incluyendo chistes y humor negro especialmente grave y en relación con eventos que, no siendo violentos (muerte natural o accidental), causan mucho dolor a víctimas indirectas y otras que afectan a sentimientos o creencias aceptadas por una colectividad.

Por lo que si el comentario leído por los etiquetadores presenta alguna de estas formas de expresión en su contenido; quedará definido como un comentario con contenido apológico.

6. Ejercicios de Prueba

En esta ultima fase del proceso de capacitación de los etiquetadores, se realizaran ejercicios de prueba en donde se les presentaran comentarios con contenido apológico y comentarios que carecen de dicho contenido. Esto con el propósito de corroborrar que los etiquetadores hayan comprendido las fases anteriores; y que al momento de llevar a cabo la tarea de etiquetación esta sea realizada de manera correcta.

En caso de que estos se equivoquen al momento de etiquetar un comentario se les presentará de nuevo las caracteristicas que deben de identificar hasta que ellos logren identificar de manera correcta cuales comentarios si estan presentando o no contenido apológico.

A continuación, se muestra en la Tabla 11 algunos de los comentarios que se utilizaran como para los ejercicios de prueba.

Tabla 11: Comentarios de prueba

No.	Comentario	Apológico	No Apológico	Indeciso
1	hoy más que nunca hay que matar a todos los chairos	x		
2	me desperté con ganas de matar varones		x	
3	Ahhh neta quiero matar al conductor con toda mi alma, lo odioooooooooooo		x	
4	Yo estoy dispuesto a matar a López Obrador y a quién o quiénes lo defiendan. Sin que me paguen, nomas por el puro placer de quitarles a su falso profeta acabar de una vez por todas con ese cáncer sexenal,	x		
5	Voy a disparar a @lopezobrador_ más de 500 veces, y si alguien se pone delante lo descuartizaré de la peor forma.	x		
6	A John Lennon lo mató un fan. A Versace lo mató un fan. A Selena la mató un fan. A ver a qué hora, chairos.	x		

Una vez realizado el Proceso de capacitación a los etiquetadores, se realizó la validación de los datos que conforman el corpus utilizando una muestra de 1000 comentarios pertenecientes al total de 4600 comentarios que conforman al corpus apológico creado, tal como se muestra en la Figura 12.

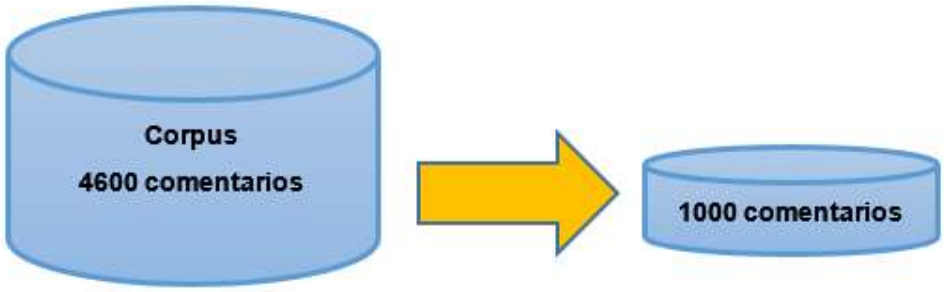


Figura 12: Muestra de comentarios a etiquetarse

Los comentarios fueron proporcionados en un archivo tipo csv, a un grupo de cinco individuos pertenecientes a los posgrados de maestría y doctorado en Ciencias de la Computación los cuales fungieron como etiquetadores

Para realizar el etiquetado de los comentarios se consideraron tres clases:

- Apológico (Comentarios con contenido apológico)
- No apológico (Comentarios sin contenido apológico)
- Indeciso (Comentarios, que los etiquetadores no saben si es apológico o no apológico)

Los etiquetadores llevaron a cabo la lectura de cada uno de los comentarios que se les proporcionaron determinando de acuerdo a su criterio personal y conocimientos a que clase consideraban que pertenecían dichos comentarios; de igual forma sustentaban su respuesta considerando los tipos de contenidos que se les fueron presentados en el proceso de capacitación de etiquetadores.

A continuación, en la Tabla 12 se observa un ejemplo de comentarios en los que los individuos coincidieron que dichos comentarios presentaban contenido apológico.

Tabla 12: Ejemplo de comentarios etiquetados como apológicos

No	***Comentario***	***Etiquetadores***				
		1	*2*	*3*	*4*	*5*
9	hoy más que nunca hay que matar a todos los chairos	A	A	A	A	A
15	Matar en México es impunidad garantizada					
34	Maten a @lopesobrador_, hay miles de formas de hacerlo sin que nadie sepa	A	A	A	A	A
28	Se dónde vives y por donde te mueves, tu y toda tu familia. Andate con cuidado	A		A	A	
40	A John Lennon lo mató un fan. A Versace lo mató un fan. A Selena la mató un fan. A ver a que hora, chairos.	A	A	A	A	A
50	Voy a decapitar tu cabeza	A	A		A	
51	Te voy a dar tantos machetazos que voy a dejarte la cada hecha trocitos y luego haré una sopa con esa carne hijo de la gran puta.	A	A	A	A	A

Los resultados obtenidos por cada uno de los individuos al momento de realizar el etiquetado de los comentarios se muestran que el 81.18% de ellos fueron etiquetados por los cinco individuos como comentarios con contenido apológico tal como se muestra en la Tabla 13.

Tabla 13: Resultados obtenidos del etiquetado

Etiquetador	**Apológico**	**No Apológico**	**Total**
E1	822	178	1000
E2	816	184	
E3	837	163	
E4	768	232	
E5	816	184	
Porcentaje	81.18	18.82	100

Con los resultados obtenidos se procedió a garantizar la concordancia entre los valores; esto mediante el coeficiente Kappa (Kohen, 1960) el cual se encarga de medir el índice de concordancia y fiabilidad entre los etiquetadores; mediante la siguiente ecuación:

$$k = \frac{Po - Pc}{1 - Pc} \qquad (4.1)$$

Donde *Po* es la proporción de unidades en la que los observadores coincidieron y *Pc* es la proporción de unidades para las cuales se espera un acuerdo en la hipótesis de independencia entre los observadores.

Para el calculó de *Po* se utiliza la siguiente ecuación

$$Po = \frac{a - d}{N} \qquad (4.2)$$

De acuerdo al caso de estudio los datos se representarían de la siguiente manera: *N* es el número total de comentarios leídos, *a* es el número de comentarios en que ambos observadores etiquetaron como Apológico a los que consideraban que presentaba este contenido y *d* es el número de comentarios que fueron etiquetados por ambos observadores como No Apológicos.

Por último, para realizar el calculó de *Pc* se utiliza la siguiente ecuación:

$$Po = \frac{rt + su}{N^2} \qquad (4.3)$$

En donde *t* y *u* son considerados como el número total de comentarios etiquetados por el observador A de acuerdo a las clases que se consideren, *r* y *s* son el número total de comentarios etiquetados por el observador B; los valores de b y c representan el número de comentarios en que los observadores no coincidieron al etiquetarlos de acuerdo a la clase a la que pertenecen.

Todo lo mencionado anteriormente se puede representar en forma tabular tal como se presenta en la Tabla 14.

Tabla 14: Cálculo de índice Kappa

Grupo N			
	Observador A		
Observador B	Apológico	No Apológico	Total
Apológico	*a*	*b*	*r*
No Apológico	*c*	*d*	*s*
Total	*t*	*u*	*N*

En donde los valores dentro de la celda sombreada representan el número de comentarios en que ambos observadores coincidieron en que dichos comentarios pertenecen a la clase que se está etiquetando.

Para llevar a cabo la comprobación de que los valores obtenidos en la Tabla 13 fueran correctos y que presentaban una alta fiabilidad en la valoración de todos y cada uno de ellos; consideró la muestra de los 1,000 comentarios previamente etiquetados y se hizo uso de los individuos que participaron en la etiquetación; con los cuales se crearon cuatro grupos, cada uno de ellos compuesto por dos individuos los cuales fueron nombrados observadores debido a que nuevamente procedieron a la lectura de los 1,000 comentarios.

Dichos individuos debían de afirmar si el mensaje presentaba alguna conducta apológica y de no ser así por alguno de los dos observadores, el comentario era etiquetado como No apológico o quedaba descartado.

Creados los cuatro grupos de análisis, se procedió con el cálculo del índice Kappa en cada uno de ellos. A continuación, en la Tabla 15 se muestra el calculó de los datos para el Grupo 1, en donde se obtuvo que en 788 comentarios ambos observadores coincidieron que presentaban contenido Apológico.

Tabla 15: Cálculo del índice Kappa al Grupo 1

Grupo 1			
	Observador 1		
Observador 2	Apológico	No Apológico	Total
Apológico	788		816
No Apológico		153	184
Total	822	178	1000

En la Tabla 16 se observa que los observadores coincidieron en que 796 son Apológicos.

Tabla 16: Cálculo del índice Kappa al Grupo 2

Grupo 2			
	Observador 2		
Observador 3	Apológico	No Apológico	Total
Apológico	796		837
No Apológico		149	163
Total	816	184	1000

En la Tabla 17 se observa que los observadores del Grupo 3 coincidieron en que 762 son comentarios que presentan contenido Apológico.

Tabla 17: Cálculo del índice Kappa al Grupo 3

Grupo 3			
	Observador 3		
Observador 4	Apológico	No Apológico	Total
Apológico	762		768
No Apológico		161	232
Total	837	163	1000

En la Tabla 18 se observa que los observadores coincidieron en que 801 son comentarios Apológicos.

Tabla 18: Cálculo del índice Kappa al Grupo 4

Grupo 4			
	Observador 4		
Observador 5	Apológico	No Apológico	Total
Apológico	801		816
No Apológico		160	184
Total	837	163	1000

El resultado obtenido de esta comprobación fue de (0.80), por lo que la fuerza de concordancia entre los etiquetadores se clasifica como “buena” según los parámetros descritos en la Tabla 19.

Tabla 19: Resultados del índice de concordancia Kappa

Grupo	Resultado
Grupo 1	0.80
Grupo 2	0.80
Grupo 3	0.75
Grupo 4	0.86
Promedio	**0.80**

4.3 Definición de las Reglas Difusas

Para el desarrollo del modelo difuso que permita el análisis de comentarios extraídos de la red social Twitter se tuvo que realizar un análisis sobre cuáles son las variables lingüísticas que permiten clasificar un comentario en Apológico y No Apológico basándose en el contenido de los mismos; una variable lingüística es aquella cuyos valores son palabras o sentencias en un lenguaje natural o artificial *(Zadeh, 1973)*

Las variables seleccionadas para la creación del conjunto de reglas difusas son: Puntuación, Incidencia y Sentimiento; las cuales son consideras como variables de

entrada que al combinarse entre sí arrojan una variable de salida la cual lleva por nombre Clase.

A continuación, se describen cada una de las variables:

- Variable Puntuación (valor numérico en el intervalo [0, 1]): Es aquella en donde su clasificación está basada en tres parámetros: Bajo, Medio y Alto, las cuales representan el valor de cada palabra clave.
- Variable Sentimiento (valor numérico en el intervalo [0, 1]): Es compuesta por cinco categorías: Miedo, Tristeza, Machismo, Discriminación y Enojo; siendo estas uno de los principales motivos en la generación de comentarios con contenido apológico.
- Variable Incidencia (valor numérico en el intervalo [0, 1]): Se compone por tres categorías: Poco frecuente, Frecuente y Muy frecuente, estás representan la aparición de palabras clave en el análisis de comentarios.
- Variable Clase (valor numérico en el intervalo [0, 1]): Se compone por dos categorías: Apológico y No Apológico, las cuales representan si un comentario presenta contenido apológico o no.

Es importante mencionar que el conjunto de reglas difusas forma parte principalmente de la Base de conocimiento del modelo difuso, debido a que permiten la transferencia de experiencia modelos, las cuales están asociadas con las funciones de pertenencia (MF).

En este trabajo de investigación se definieron 15 reglas de tipo Takagi-Sugeno y son las que se muestran a continuación; representando por color **azul** las variables de entrada, por color **marrón** los valores de las variables y por color **verde** las variables de salida.

R^1: Si Puntuacion Es Bajo Y Sentimiento Es Tristeza Y Incidencia Es Poco_Frecuente ENTONCES Clase Es No_Apologico

R^2: Si Puntuacion Es Bajo Y Sentimiento Es Miedo Y Incidencia Es Poco_Frecuente ENTONCES Clase Es No_Apologico

R^3: Si Puntuacion Es Bajo Y Sentimiento Es Machismo Y Incidencia Es Poco_Frecuente ENTONCES Clase Es Apologico

R^4: Si Puntuacion Es Bajo Y Sentimiento Es Discriminación Y Incidencia Es Poco_Frecuente ENTONCES Clase Es Apologico

R^5: Si Puntuacion Es Bajo Y Sentimiento Es Enojo Y Incidencia Es Poco_Frecuente ENTONCES Clase Es Apologico

R^6: Si Puntuacion Es Medio Y Sentimiento Es Tristeza Y Incidencia Es Frecuente ENTONCES Clase Es No_Apologico

R^7: Si Puntuacion Es Medio Y Sentimiento Es Miedo Y Incidencia Es Frecuente ENTONCES Clase Es No_Apológico

R^8: Si Puntuacion Es Medio Y Sentimiento Es Machismo Y Incidencia Es Frecuente ENTONCES Clase Es Apológico

R^9: Si Puntuacion Es Medio Y Sentimiento Es Discriminación Y Incidencia Es Frecuente ENTONCES Clase Es Apologico

R^{10}: Si Puntuacion Es Medio Y Sentimiento Es Enojo Y Incidencia Es Frecuente ENTONCES Clase Es Apologico

R^{11}: Si Puntuacion Es Alto Y Sentimiento Es Tristeza Y Incidencia Es Muy_Frecuente ENTONCES Clase Es No_Apologico

R^{12}: Si Puntuacion Es Alto Y Sentimiento Es Miedo Y Incidencia Es Muy_Frecuente ENTONCES Clase Es No_Apologico

R^{13}: Si Puntuacion Es Alto Y Sentimiento Es Machismo Y Incidencia Es Muy_Frecuente ENTONCES Clase Es Apologico

R^{14}: Si Puntuacion Es Alto Y Sentimiento Es Discriminación Y Incidencia Es Muy_Frecuente ENTONCES Clase Es Apologico

R^{15}: Si Puntuacion Es Alto Y Sentimiento Es Enojo Y Incidencia Es Muy_Frecuente ENTONCES Clase Es Apologico

4.3.1 Representación de las variables en funciones de pertenencia

La selección de las funciones de pertenencia a utilizarse en este trabajo de investigación, se llevó mediante el análisis de los datos con los que se estaría trabajando; optando por trabajar con funciones de pertenencia tipo Triangulares y Trapezoidales, las cuales debido a la magnitud de los datos a analizar nos permiten realizar de una manera más simple la identificación de los datos; las funciones de pertenencia nos permiten representar valores del mundo real para que el modelo difuso los analice y realice ciertas acciones de acuerdo a los valores asignados o definidos en un rango establecido.

A continuación, se muestran las funciones de pertenencia utilizadas:

Variable de entrada Puntuacion

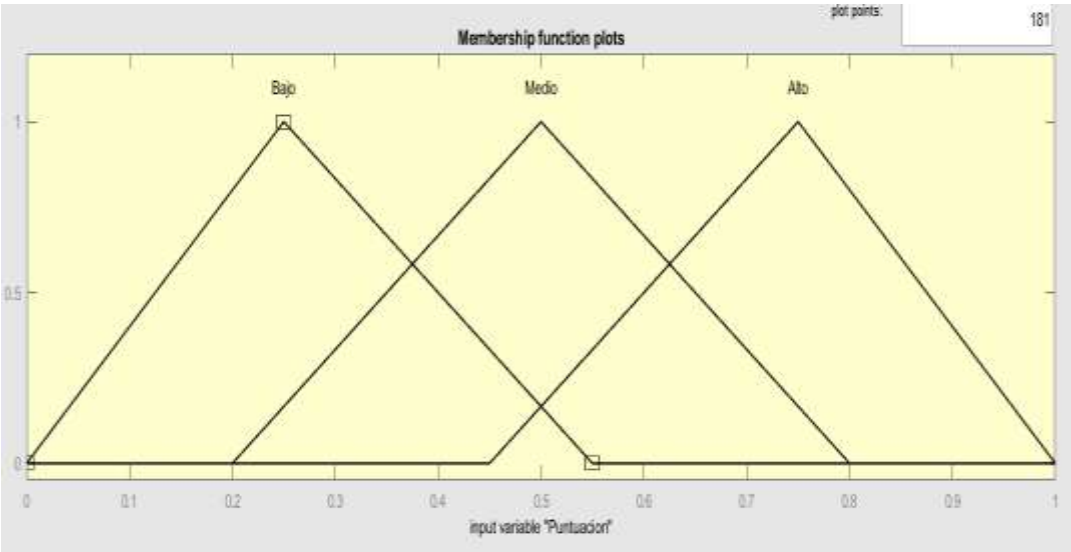


Figura 13 Funciones de pertenencia de la variable Puntuacion

En la Figura 13 la variable de entrada Puntuacion se representa entre el rango de valores [0,1] mediante una función de pertenencia triangular, compuesta por tres categorías las cuales se encuentran entre los siguientes valores:

- Bajo (0, 0.25, 0.55)
- Medio (0.20, 0.50, 0.80)
- Alto (0.45, 0.75, 1)

Cada una de las funciones que conforman la variable de Puntuación cuentan con un valor de radio de 0.3 entre el núcleo de cada una de las funciones la frontera de las mismas.

Variable de entrada Sentimiento

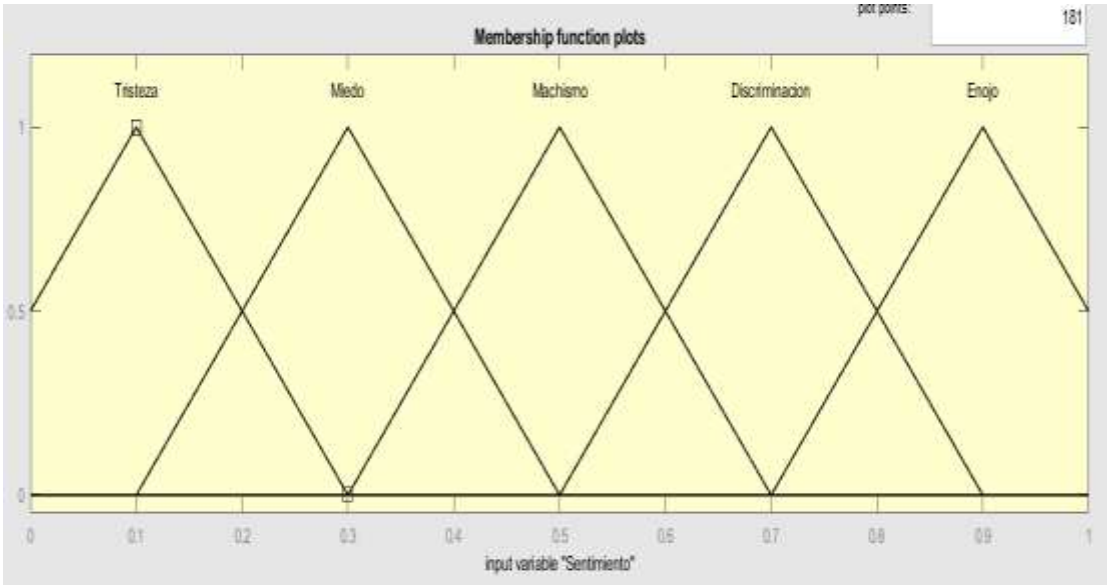


Figura 14: Funciones de pertenencia de la variable Sentimiento

La variable de entrada Sentimiento se representa entre el rango de valores [0,1] (Figura 14) mediante una función de pertenencia triangular, compuesta por cinco categorías las cuales se encuentran entre los siguientes valores:

- Tristeza (-0.1, 0.1, 0.3)
- Miedo (0.1, 0.3, 0.5)
- Machismo (0.3, 0.5, 0.7)
- Discriminación (0.5, 0.7, 0.9)
- Enojo (0.7, 0.9, 1.1)

Las cuales cuentan con un valor de radio de 0.2 entre el núcleo le función de pertenencia y la frontera de la misma.

Variable de entrada Incidencia

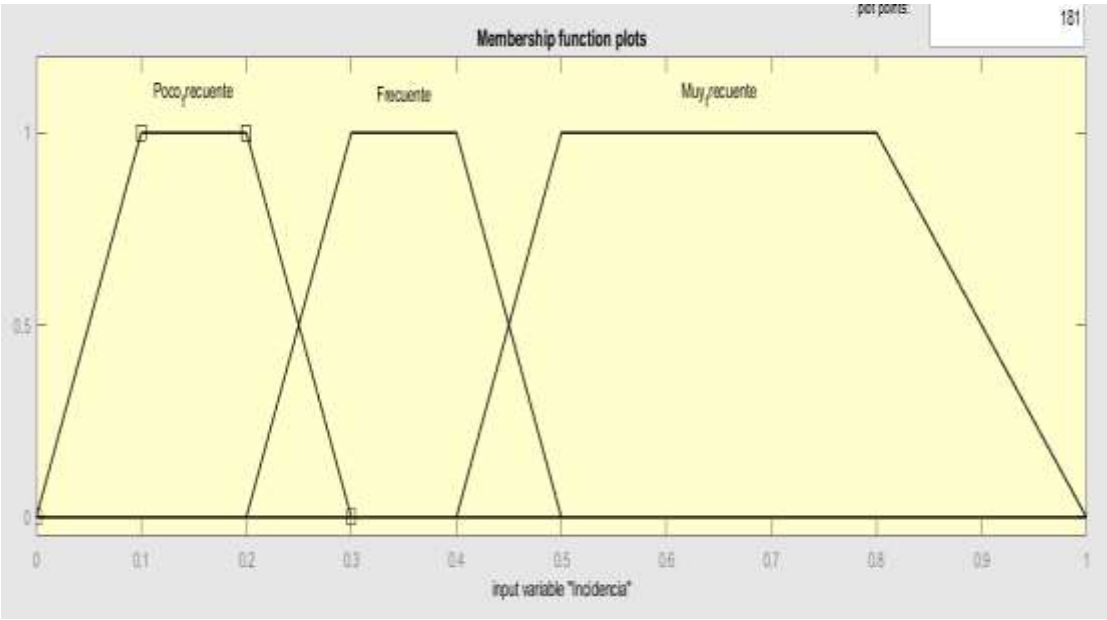


Figura 15: Funciones de pertenencia de la variable Incidencia

La variable de entrada Incidencia se representa entre el rango de valores [0,1] (Figura 15) mediante una función de pertenencia trapezoidal compuesta por tres categorías las cuales se encuentran entre los siguientes valores:

- Poco_frecuente (0.0, 0.1, 0.2, 0.3)
- Frecuente (0.2, 0.3, 0.4, 0.5)
- Muy_frecuente (0.4, 0.5, 0.8, 1.0)

Variable de salida Clase

La variable de salida Clase se representa entre el rango de valores [0, 1], mediante una representación binaria compuestas por los siguientes valores:

- Apológico (0.5, 1), es considerado como comentarios Apológico cuando el valor final es mayor que 0.5
- No_Apológico (0, 0.5), es considerado como comentario No_apológico cuando el valor obtenido es menor que 0.5.

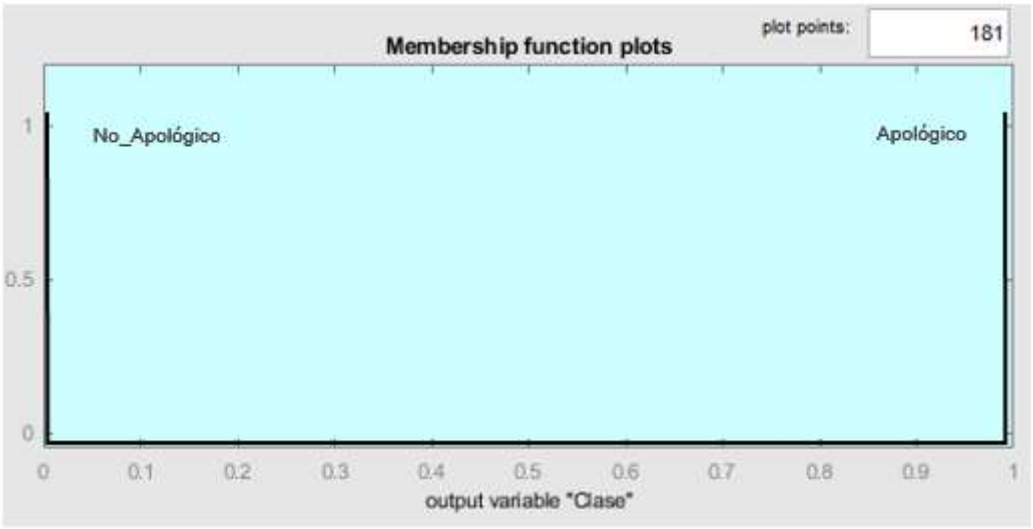


Figura 16: Funciones de pertenencia de la variable Clase

4.3.2 Estructura del modelo difuso

Un modelo difuso es desarrollado con el propósito de poder solucionar situaciones de incertidumbre, mediante la representación del conocimiento humano por medio de reglas de inferencia difusa que nos permiten asignar variables lingüísticas de entrada y salida para analizar nuestro conjunto de datos.

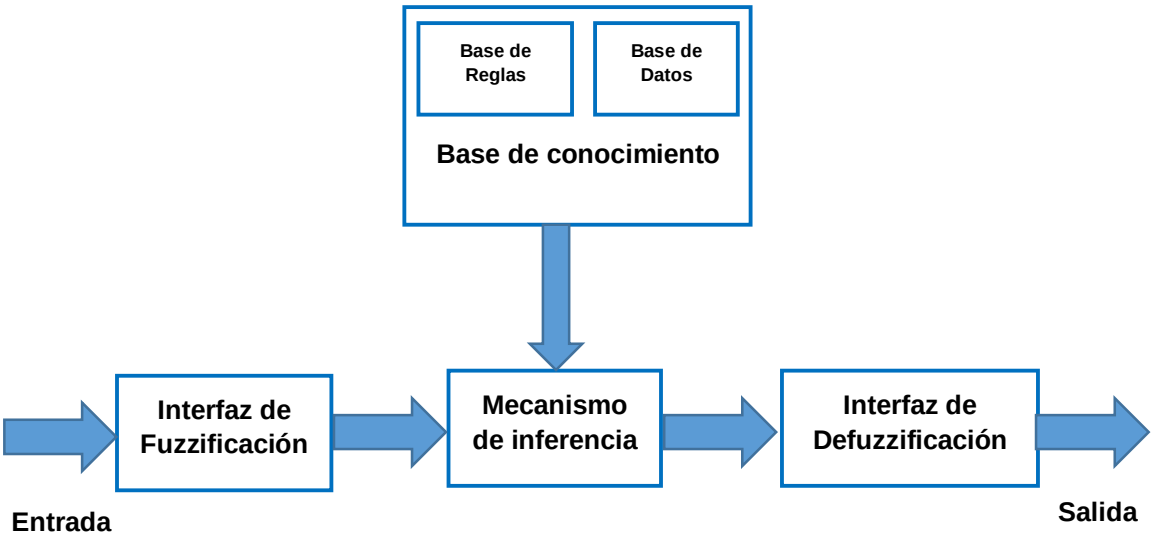


Figura 17 Estructura del Sistema Basado en Reglas Difusas

La estructura del modelo difuso desarrollado se puede observar en la *Figura 17;* dicho modelo se compone por cinco módulos los cuales son: Interfaz de Fuzzificación, Mecanismo de inferencia, Base de conocimiento e Interfaz de Defuzzificación a continuación se describe cada uno de los módulos.

- El módulo de Fuzzificación es el encargado de llevar a cabo el proceso de agrupar un valor numérico a un dominio difuso, permitiendo así asignarle un valor de pertenecía.
- La Base de Conocimiento (BC): es aquella que se compone principalmente por las reglas difusas establecidas y las funciones de pertenecía que representan a cada una de las variables previamente definidas. Para este trabajo de investigación se hizo uso de tres variables de entrada y una variable de salida; las cuales representan el conocimiento que se tiene sobre el problema que vamos resolver.
- El mecanismo de inferencia difusa es en donde se aplican las reglas establecidas en la base de conocimiento a fin analizar los datos y de determinar una salida.
- En el módulo de Defuzzificación se llevaba cabo el proceso inverso de la Fuzzificación, es este caso se asigna un valor numérico a un valor que se encontraba en el dominio difuso, en este trabajo de investigación la salida del modelo es binaria es decir 0 o 1, dependiendo de los grados de pertenencia que se obtengan en el mecanismos de inferencia; es decir si el grado de pertenencia es mayor o igual de 0.5 se obtendrá como salida 1 que corresponde a comentario apológico y si el grado de pertenencia es menor que 0.5 el valor de salida es 0.

Capítulo 5 Pruebas y Resultados

En este capítulo se presentan los experimentos diseñados e implementados para este trabajo, así como las métricas de evaluación y resultados obtenidos.

5.1 Diseño e Implementación de Pruebas

En esta sección se detallan los experimentos y métricas implementadas al modelo difuso.

5.1.1 Métricas

Para validar que los datos obtenidos con el sistema de clasificación son correctos se hizo uso de la matriz de confusión y sus métricas correspondientes; la cual es una matriz de tipo n x n en la que las filas se nombran según las clases reales y las columnas, según previstas por el modelo. Por lo que dicha matriz sirve para mostrar de forma explícita cuándo una clase es confundida con otra; la estructura de dicha matriz se muestra en la Figura 18.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 18: Matriz de confusión (RPubs, 2017)

Donde:

- **VP:** es la cantidad de datos que fueron clasificados correctamente como positivos por el modelo; un comentario etiquetado correctamente como Apológico.
- **VN:** es la cantidad de datos que fueron clasificados correctamente como negativos por el modelo; un comentario etiquetado correctamente como No apológico.

- **FN:** es la cantidad de datos que fueron clasificados incorrectamente como negativos; es decir un comentario etiquetado como No apológico pero que en realidad si es Apológico.
- **FP** es la cantidad de datos que fueron clasificados incorrectamente como positivos; es decir como un comentario etiquetado como Apológico pero que en realidad es No apológico.

Esta matriz se puede representar también de forma gráfica la cual es más entendible para el usuario; tal como se observa en la Figura 19.

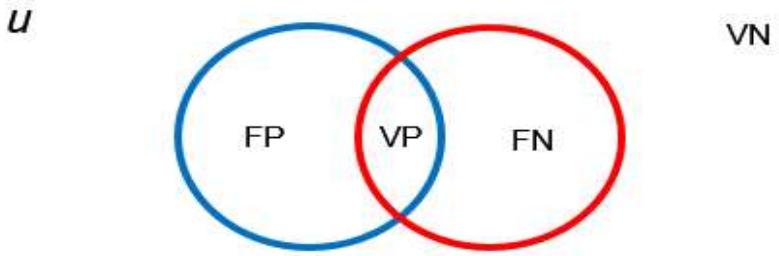


Figura 19: Representación en conjuntos de la matriz de confusión (Datasource, 2020)

Donde:

Toda el área cubierta por el recuadro negro representa el universo de datos (*u*)

El área dentro del círculo azul representa los datos de estudio, es decir aquellos obtenidos por el sistema.

El área dentro del círculo rojo representa los datos objetivos, es decir aquellos determinados por los expertos.

Para la evaluación del modelo se utilizaron las métricas de Precisión, Cobertura y F1-Score. A continuación, se describe cada una de estas métricas con su respectiva ecuación:

Precisión: Representa que fracción de los datos que el sistema cree que son correctos, son realmente correctos o se aproximan a los correctos y se define como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos positivos:

$$Precisión = \frac{VP}{FP + VP} \quad (5.1)$$

Cobertura: Representa que fracción de los datos selecciona como correctos el sistema y se define como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos.

$$Cobertura = \frac{VP}{VP + FN} \quad (5.2)$$

F1-Score: Combina las medidas de precisión y cobertura para devolver una medida de calidad más general del modelo se le conoce como la media armónica de las métricas mencionadas, donde entre más se acerca el valor a 1 es el mejor valor (es decir perfecta Precisión y Cobertura) y peor si se acerca a 0. Para interpretar los valores obtenidos por las métricas, en 5.3 se muestra la composición de esta media.

$$F1 = 2 * \frac{Precisión * Cobertura}{Precisión + Cobertura} \quad (5.3)$$

5.2 Pruebas

Una vez determinados dichos conceptos se realizó el análisis de la precisión que tiene el sistema difuso para etiquetar comentarios; para esto se hizo uso de los datos recabados en semestres pasados el cual corresponde a un corpus de 4,600 comentarios el cual también fue analizado por el modelo difuso; con el objetivo de obtener la precisión que tiene el modelo al momento de llevar a cabo la clasificación de los comentarios de acuerdo al contenido identificados en ellos.

Se fracciono el corpus en cinco conjuntos de datos dado que las características del equipo impedían aplicar el modelo a conjunto de datos con una mayor cantidad de datos (Memoria RAM de 12 Gb, Procesador AMD A9-9410 RADEON R5).

Dichos datos también fueron clasificados de manera manual por los analistas de contenido designados para la creación del corpus; esto con el fin de poder observar el comportamiento de dicho modelo evaluándolo con las métricas de Precisión,

Cobertura y F1-Store. A continuación, se describen cada una de las pruebas realizadas.

5.2.1 Prueba 1

El archivo utilizado está compuesto por 1500 comentarios de los cuales 700 presentan contenido apológico y 800 contenido no apológico.

Al realizar el etiquetado del mismo archivo con el modelo basado en reglas difusas, se obtuvo como resultado que 540 comentarios presentan contenido apológico y 960 contenido no apológico; tal como se muestra en la Tabla 20.

Tabla 20: Matriz de confusión de la Prueba 1

		Predicción		**Total**
		Positivos	*Negativos*	
Observación	*Positivo*	(VP) 483	(FN) 217	700
	Negativo	(FP) 57	(VN) 743	800
	Total	540	960	1500

En la Figura 20 se muestra la representación de los datos etiquetados por el sistema difuso.

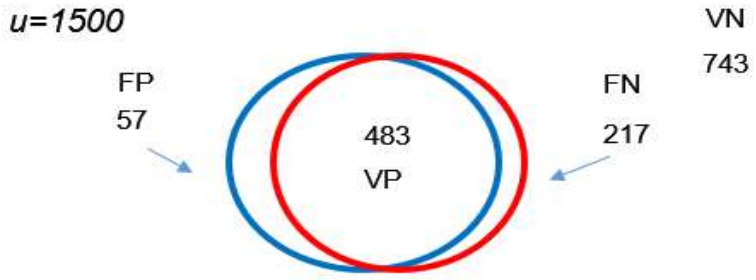


Figura 20: Representación de los datos de la Prueba 1

Una vez presentados los datos se prosiguió a realizar el cálculo de las métricas de evaluación tal como se muestra en la Tabla 21.

Tabla 21: Métricas de evaluación para la Prueba 1

Cálculo de Precisión	Cálculo de Cobertura	Cálculo de F1-Store
$P = \frac{483}{57+483} = \frac{483}{540} = 0.89$	$C = \frac{483}{483+217} = \frac{483}{700} = 0.69$	$F1 = 2 * \frac{0.89 * 0.69}{0.89 + 0.69} = 2 * \frac{0.61}{1.58} = 2 * 0.38 = 0.76$

5.2.2 Prueba 2

El archivo que se utilizó está compuesto por 850 comentarios de los cuales 383 presentan contenido apológico y 467 contenido no apológico.

Al realizar el etiquetado del mismo archivo con el sistema basado en reglas difusas se obtuvo como resultado que 359 comentarios presentan contenido apológico y 491 comentarios contenido no apológico; tal como se muestra en la Tabla 22.

Tabla 22: Matriz de confusión de la Prueba 2

		Predicción		**Total**
		Positivos	*Negativos*	
Observación	*Positivo*	(VP) 260	(FN) 123	383
	Negativo	(FP) 99	(VN) 368	467
	Total	359	491	850

En la Figura 21 se muestra la representación de los datos etiquetados por el sistema difuso.

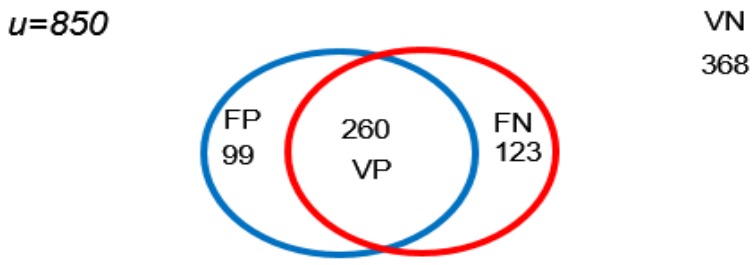


Figura 21: Representación de los datos de la Prueba 2

Una vez presentados los datos se prosiguió a realizar el cálculo de las métricas de evaluación tal como se muestra en la Tabla 23.

Tabla 23: Métricas de evaluación para la Prueba 2

Cálculo de Precisión	Cálculo de Cobertura	Cálculo de F1-Store
$P = \frac{260}{99+260} = \frac{260}{359} = 0.72$	$C = \frac{260}{260+123} = \frac{260}{383} = 0.67$	$F1 = 2 * \frac{0.72 * 0.67}{0.72 + 0.67} = 2 * \frac{0.48}{1.39} = 2 * 0.34 = 0.68$

5.2.3 Prueba 3

El archivo que se utilizó está compuesto por 1250 comentarios de los cuales 927 presentan contenido apológico y 323 contenido no apológico.

Al realizar el etiquetado del mismo archivo con el sistema basado en reglas difusas se obtuvo como resultado que 796 comentarios presentan contenido apológico y 454 comentarios contenido no apológico; tal como se muestra en la Tabla 24.

Tabla 24: Matriz de confusión de la Prueba 3

		Predicción		**Total**
		Positivos	*Negativos*	
Observación	*Positivo*	(VP) 612	(FN) 315	927
	Negativo	(FP) 184	(VN) 139	323
	Total	796	454	1250

En la Figura 22 se muestra la representación de los datos etiquetados por el sistema difuso.

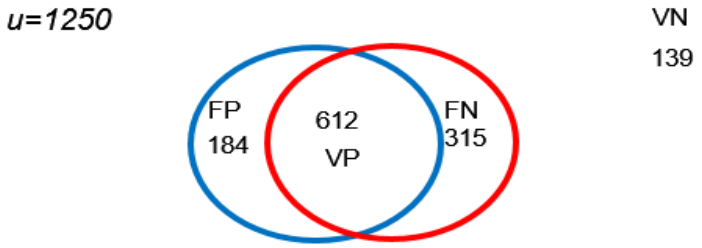


Figura 22: Representación de los datos de la Prueba 3

Una vez presentados los datos se prosiguió a realizar el cálculo de las métricas de evaluación tal como se muestra en la Tabla 25.

Tabla 25 Métricas de evaluación para la Prueba 3

Cálculo de Precisión	Cálculo de Cobertura	Cálculo de F1-Store
$P = \frac{612}{184 + 612} = \frac{612}{796} = 0.76$	$C = \frac{612}{612 + 315} = \frac{612}{927} = 0.66$	$F1 = 2 * \frac{0.76 * 0.66}{0.76 + 0.66} = 2 * \frac{0.50}{1.42} = 2 * 0.35 = 0.70$

5.2.4 Prueba 4

El archivo que se utilizó está compuesto por 300 comentarios de los cuales 246 presentan contenido apológico y 54 contenido no apológico.

Al realizar el etiquetado del mismo archivo con el sistema basado en reglas difusas se obtuvo como resultado que 221 comentarios presentan contenido apológico y 79 comentarios contenido no apológico; tal como se muestra en la Tabla 26.

Tabla 26: Matriz de confusión de la Prueba 4

		Predicción		**Total**
		Positivos	*Negativos*	
Observación	*Positivo*	(VP) 175	(FN) 70	246
	Negativo	(FP) 45	(VN) 9	54
	Total	221	79	850

En la Figura 23 se muestra la representación de los datos etiquetados por el sistema difuso.

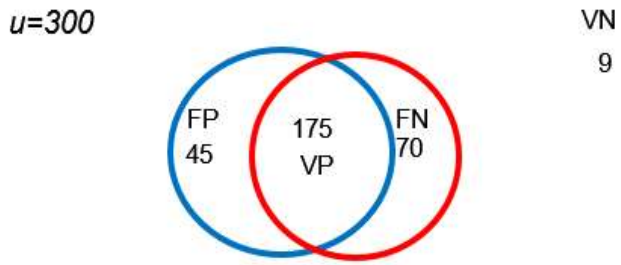


Figura 23: Representación de los datos de la Prueba 4

Una vez presentados los datos se prosiguió a realizar el cálculo de las métricas de evaluación tal como se muestra en la Tabla 27.

Tabla 27: Métricas de evaluación para la Prueba 4

Cálculo de Precisión	Cálculo de Cobertura	Cálculo de F1-Store
$P = \frac{175}{45 + 175} = \frac{175}{220} = 0.79$	$C = \frac{175}{175 + 70} = \frac{175}{245} = 0.71$	$F1 = 2 * \frac{0.79 * 0.71}{0.79 + 0.71}$ $= 2 * \frac{0.56}{1.5}$ $= 2 * 0.37$ $= 0.74$

5.2.5 Prueba 5

El archivo que se utilizó está compuesto por 2600 comentarios de los cuales 1938 presentan contenido apológico y 662 contenido no apológico.

Al realizar el etiquetado del mismo archivo con el sistema basado en reglas difusas se obtuvo como resultado que 1781 comentarios presentan contenido apológico y 819 comentarios contenido no apológico; tal como se muestra en la Tabla 28.

Tabla 28: Matriz de confusión de la Prueba 5

		Predicción		**Total**
		Positivos	*Negativos*	
Observación	*Positivo*	(VP) 1583	(FN) 355	1938
	Negativo	(FP) 198	(VN) 464	662
	Total	1781	819	2600

En la Figura 24 se muestra la representación de los datos etiquetados por el sistema difuso.

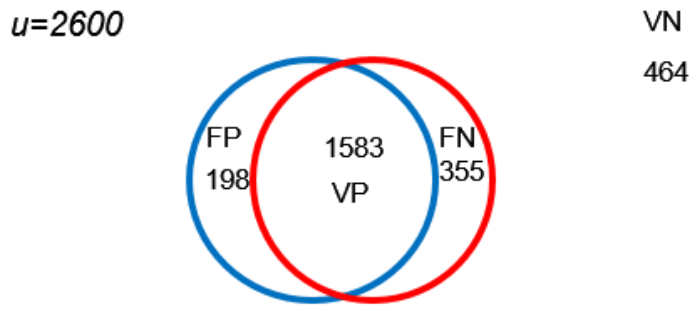


Figura 24: Representación de los datos de la Prueba 5

Una vez presentados los datos se prosiguió a realizar el cálculo de las métricas de evaluación tal como se muestra en la Tabla 29.

Tabla 29: Métricas de evaluación para la Prueba 5

Cálculo de Precisión	Cálculo de Cobertura	Cálculo de F1-Store
$P = \frac{1583}{198 + 1583} = \frac{1583}{1781} = 0.88$	$C = \frac{1583}{1583 + 355} = \frac{1583}{1938} = 0.81$	$F1 = 2 * \frac{0.88 * 0.81}{0.88 + 0.81} = 2 * \frac{0.71}{1.69} = 2 * 0.42 = 0.84$

5.3 Resultados

A continuación en la Tabla 30, se muestran los resultados obtenidos de las pruebas realizadas.

Tabla 30: Resultados obtenidos del modelo difuso

No. de Prueba	No. de Comentarios	Valores Reales		Valores Modelo Difuso		Métricas de Evaluación		
		Apológico	No Apológico	Apológico	No Apológico	Precisión	Cobertura	F1-Store
1	1500	700	800	540	960	0.89	0.69	0.76
2	850	383	467	359	491	0.72	0.67	0.68
3	1250	927	323	796	454	0.76	0.66	0.70
4	300	246	54	221	79	0.79	0.71	0.74
5	2600	1938	662	1781	819	0.88	0.81	0.84

Al momento de llevar el análisis de los datos obtenidos de cada una de las pruebas podemos observar que en la Prueba 1 se utilizó un conjunto de 1500 comentarios de los cuales el modelo difuso etiqueto correctamente un total de 487 comentarios con contenido apológico; obteniendo una precisión de 89% la cual nos dice que los datos que el modelo cree que son correctos, efectivamente lo son.

En la Prueba 2 se utilizó un total de 850 comentarios, en donde el sistema difuso etiqueto como apológicos un total de 359 comentarios de los cuales 260 si efectivamente presentan contenido apológico; por lo que en dicha prueba el valor de la precisión del modelo fue de un 72% el cuál es el valor más bajo obtenido de las cinco pruebas realizadas.

La Prueba 3 se realizó con un total de 1250 comentarios, de los cuales el modelo difuso etiqueto un total de 796 comentarios como apológicos; acertando correctamente en 612, por lo que el resto de los comentarios no presentan contenido apológico.

En Prueba 4 se analizó una pequeña muestra de comentarios compuesta por un total de 300 comentarios; de los cuales el modelo etiqueto como apológicos 221 y 79 como no apológicos. Sin embargo, solo coincidió en que 179 presentan contenido apológico y 9 no presentan contenido apológico.

La última de las pruebas realizadas se llevó a cabo con un conjunto de datos de 2600 comentarios; siendo la muestra de datos más representativa; en dicha prueba el modelo difuso etiqueto 1781 comentarios como apológicos y 819 como no apológicos; de los cuales 1583 efectivamente si son apológicos; por lo que podemos observar que el margen de error en el etiquetado de datos es bajo, aunque no debemos de pasarlo por alto. En esta última prueba se obtuvo como resultado el 88% de precisión y el 81% de cobertura.

Es importante mencionar que, al momento de llevar a cabo la evaluación de los resultados obtenidos con el sistema, se debe de verificar que los datos se clasifiquen correctamente en las clases establecidas.

Los resultados obtenidos de las pruebas realizadas con el Sistema difuso nos demuestran que independientemente del número de comentarios a analizarse, los

valores de precisión se encuentran por arriba del 70%, considerándose como una precisión buena, la igual que el valor de F1-Store donde se sabe que mientras el valor se acerque más al valor de 1 quiere decir que es el mejor valor.

Capítulo 6 Conclusiones

Con el desarrollo de este trabajo de investigación se logró crear un método basado en lógica difusa que permite realizar una automatización del proceso de análisis que conlleva el analizar comentarios extraídos de la red social Twitter, con la finalidad de apoyar en la identificación de comentarios que presentan Apologías del Delito es decir algún tipo de incitación al odio o violencia.

De igual forma logró adquirir comentarios de la red social Twitter mediante el desarrollo de un algoritmo automático de descarga enfocado principalmente a utilizar el conjunto de palabras clave definidas en este trabajo de investigación.

Es importante mencionar que debido a que en investigaciones realizadas en la literatura no se encontraron conjuntos de textos relacionados con Apologías del delito, fue necesario llevar a cabo la creación de un Corpus Apológico compuesto por comentarios generados de usuarios que pertenecen a la red social Twitter, actualmente el corpus contiene 4600 comentarios correspondientes a las clases de Odio y Violencia a fin de poder tener un conjunto de datos relacionados con conductas apológicas que puede ser utilizados para futuras investigaciones.

Uno de los principales retos del desarrollo de este trabajo de tesis, fue seleccionar a cinco individuos que cubrieran con el perfil de un analista de contenido web para llevar a cabo la capacitación de los mismos en cuanto a tareas relacionadas con el análisis de comentarios e identificación de contenido apológico en comentarios generados por usuarios de Twitter.

Dicha capacitación se basó en el trabajo de “Taxonomía de la comunicación violenta y el discurso del odio en Internet” (Llinares, 2016), en donde se establecen los tipos de contenido que deben de presentarse comentarios relacionados con apologías del delito.

Por otro lado, en este trabajo de investigación queda demostrado que el uso de Lenguaje Natural y Lógica Difusa nos permiten automatizar tareas relacionadas con el análisis de textos haciendo uso de la inferencia difusa para identificar valores

subjetivos y poder transformarlos a valores continuos asignándoles grados de pertenencia a dichos valores, los cuales forman parte de a las funciones de pertenencia y las reglas difusas que en conjunto se le conoce como Base de conocimiento, la cual permite realizar una clasificación de dichos textos de acuerdo a las clases o categorías establecidas permitiendo reducir el tiempo en que dichas tareas se ejecutaban de forma manual.

Sin embargo, es importante mencionar que existen casos en que el modelo no es capaz de detectar si un comentario presenta contenido apológico o no y esto es debido a la ambigüedad de la lengua, ya que pueden presentarse comentarios que utilizan palabras que no están dentro del conjunto de palabras clave establecidas; por lo que sería interesante desarrollar un conjunto de datos en donde se encuentren los sinónimos de las palabras que tienen que ver con conductas delictivas al igual que considerar aquellas que mediante el sarcasmo incitan a comentar algún tipo de delito, considerando también llevar a cabo un análisis semántico del texto para poder interpretar su significado.

6.1 Trabajos futuros

Como trabajo futuro se contempla el incremento del Corpus Apológico para llevar a cabo la validación de las reglas creadas para la base de conocimiento, las cuales conforman al modelo difuso a fin de poder incrementar los valores de precisión que se obtuvieron que en promedio son del 78% y de cobertura en promedio son del 75%.

Referencias Bibliográficas

Aguado de Cea, G., Barrios, M., Bernardos, M., Campanella, I., Montiel-Ponsoda, E., Muñoz-García, O., & Rodríguez, V. (2012). Análisis de sentimientos de un corpus de redes sociales. 31er Congreso Asociación Española de Lingüística Aplicada. "Comunicación, Cognición y Cibernética". San Cristóbal de la Laguna, Tenerife.

Al-Ajlan, M., & Ykhlef, M. (Abril de 2018). Optimización de la detección de ciberacoso en Twitter basada en el aprendizaje profundo. *En 2018, 21a Conferencia Nacional de Computación de la Saudi Computer Society (NCC)*, pp. 1-5. IEEE.

Alami, S., & Elbeqqali, O. (Octubre de 2015). Creación de perfiles de Ciberdelito: técnicas de minería de textos para detectar y predecir actividades delictivas en publicaciones de microblogs. *En 2015, 10a Congreso Internacional de Sistemas Inteligentes: Teorías y Aplicaciones (SITA)*, pp. 1-5. IEEE.

Alghofaili, H., & Almishari, M. (Abril de 2018). Contrarrestar la incitación al terrorismo de los perfiles de Twitter en el contexto árabe. *En 2018, la 21a Conferencia Nacional de Computación de la Sociedad de Computación Saudita (NCC)*, pp. 1-5. IEEE.

Altay, E. y. (2018). Detección del ciberacoso en redes sociales mediante métodos de aprendizaje automático. *En 2018 Congreso Internacional sobre Big Data, Aprendizaje Profundo y Lucha contra el Ciber Terrorismo (IBIGDELFT)*, pp. 87-91. IEEE.

Arreerard, R., & Senivongse, T. (Julio de 2018). Clasificación del texto difamatorio tailandés en las redes sociales. *IEEE / ACIS, 3a Conferencia Internacional sobre Big Data, Computación en la Nube, Ciencia de Datos e Ingeniería*, pp. 73- 78. IEEE.

Baca, Y. (2014). Desarrollo de un servicio web para determinar la polaridad de textos de redes sociales en español (Tesis de Maestría). *Centro Nacional de Investigación y Desarrollo Tecnológico*, pp. 52-63.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (Abril de 2017). Aprendizaje profundo para la detección del odio en los tweets. *En los procedimientos de la 26ª Conferencia Internacional sobre World Wide Web Companion*, pp. 759-760. Comité Directivo Internacional de Conferencias World Wide Web.

Bairagi, V., & Tapaswi, N. (Marzo de 2016). Clasificación de comentarios de redes sociales utilizando una técnica clasificadora basada en difusos. *En 2016, Simposio sobre Análisis de Datos Colosales y Redes (CDAN)*, pp. 1-7. IEEE.

Barn, R., & Barn, B. (2016). An ontological representation of a taxonomy for cybercrime.

Brown, R., Walters, M., & Paterson, J. (12 de Enero de 2018). *Cómo los delitos de odio afectan a toda una comunidad*. Recuperado el 13 de Noviembre de 2018, de BBC NEWS: https://www.bbc.com/news/uk-42622767

Camejo, Y., Rondón, A., Ortiz, A. A., & Pernía, L. A. (2010). *Monografias.com*. Recuperado el 03 de Noviembre de 2018, de Procesamiento del lenguaje natural para recuperar información: https://www.monografias.com/trabajos81/procesamiento-lenguaje-natural-recuperar-informacion/procesamiento-lenguaje-natural-recuperar-informacion2.shtml

Campuzano, C. (2018). *Apología de delito*. Recuperado el 10 de Noviembre de 2018, de Mexico Enciclopedia Jurídica Online: https://mexico.leyderecho.org/apologia-del-delito/

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. *International journal of technology enhancements and emerging engineering research, 1*, 131-134. Consejo de Europa. (2001). *Marco Común Europeo de Referencia para las lenguas: aprendizaje, enseñanza, evaluación.* Recuperado el 20 de Octubre de 2020, de Consejo de Europa: https://rm.coe.int/1680459f97

Datasource, (2020). Métricas De Evaluación De Modelos En El Aprendizaje Automático. Recuperado el 13 de Marzo del 2020. https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Dewan, P., & Kumaraguru, P. (Julio de 2015). Identificación automática en tiempo real de publicaciones maliciosas en Facebook. *En 2015, 8a Conferencia Anual sobre Privacidad, Seguridad y Confianza (PST)*, pp. 85-92. IEEE.

Diccionario Jurídico Mexicano. (2007). Diccionario Jurídico Mexicano. Recuperado el Septiembre 24, 2018, de Diccionario Jurídico Mexicano Sitio web: http://diccionariojuridico.mx/definicion/apologia-del-delito/#:~:text=I.-,Apolog%C3%ADa%20proviene%20del%20lat%C3%ADn%20apologia%20y%20significa%20%E2%80%9Cdiscurso%20en%20defensa,quebrantamiento%20grave%20de%20la%20ley.

Elyezjy, N. T., & Elhaless, A. M. (2015). Investigando crímenes usando minería de textos y análisis de redes. *Revista Internacional de Aplicaciones Informáticas*, pp. 19-25.

Facchin, J. (14 de Enero de 2018). *El Blog de José Facchin.* Recuperado el 23 de Septiembre de 2018, de https://josefacchin.com/lista-redes-sociales-mas-importantes-del-planeta/

Filatova, E. (2012). Ironía y sarcasmo: generación y análisis de corpus utilizando crowdsourcing. *En Lrec*, pp. 392-398.

Frenda, S., & Banerjee, S. (2018). Análisis profundo en tweets agresivos mexicanos. *En el Tercer Taller de Evaluación de las Tecnologías del Lenguaje Humano para las Lenguas Ibéricas (IberEval 2018), Vol. 2150*, pp. 108-113. Ceur Workshop Proceedings.

García, I. P. (2004). Ensayo de un sistema de extracción de información (técnica de inteligencia artificial) en un centro de información especializado en sanidad vegetal. *Biblios*, *5*(20), 14-28.

Gómez, D. R. (2005). La lematización en español: una aplicación para la recuperación de información. *Gijón: Trea*.

Gupta, S., & Kumar, S. (2015). Detección y prevención de delitos mediante análisis de redes sociales. *Revista internacional de aplicaciones informáticas, 126*(6).

Han, B., Cook, P., & Baldwin, T. (2014). Predicción de geolocalización de usuarios de Twitter basada en texto. *Revista de Investigación en Inteligencia Artificial, Vol. 49*, pp. 451–500.

Hernández, M. B., & Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica, Vol. 32*(1), pp. 87–96.

Huang, Y. Y., Li, C. T., & Jeng, S. K. (Octubre de 2015). Redes sociales basadas en la ubicación minera para la predicción de actividad criminal. *En 2015, 24ª Conferencia de Comunicación Inalámbrica y Óptica (WOCC)*, pp. 185-189. IEEE.

Interpol. (2018). *Conectando a la policía para un mundo más seguro.* Recuperado el 19 de Noviembre de 2018, de Interpol: https://www.interpol.int/Crime-areas/Cybercrime/Cybercrime

Jaconelli, J. (2018). Incitación: un estudio sobre el crimen lingüistico. (Springer, Ed.) *Derecho penal y filosofía, Vol. 12*(2), pp. 245–265.

Jayaweera, I., & Wijayasiri, A. (2015). Análisis de delitos: análisis de delitos a través de artículos periodísticos. pp. 277-282.

Kohen, J. (1960). Un coeficiente de acuerdo a las escalas nominales. *Educational and psychological measurement, Vol. 20*(1), pp. 37-46.

Landaeta, G. (2014). *SEO para Google*. Recuperado el 15 de Septiembre de 2019, de http://googleseo.marketing/seo-que-son-stop-words-palabras-vacias/

Llinares, F. M. (2016). Taxonomía de la comunicación violenta y el discurso del odio en Internet. *IDP. Revista de Internet, Derecho y Política,* (22), pp. 82-107.

Manning, C. D., & Raghavan, P. (2008). *Introducción a la recuperación de información* (Vol. 1). Cambridge: Cambriedge University Press.

Martínez, G. (2015). *Introduccion a los corpús.* Instituto de Ingeniería, UNAM.

Mendel, J. M. (2001). Uncertain rule-based fuzzy logic systems: introduction and new directions. Prentice Hall PTR Upper Saddle River.

Mishra, A., & Vishwakarma, S. (2015). Analysis of TF-IDF Model and its Variant for Document Retrieval. *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 772-776.

Mohammad, S. (2017). Léxico de intensidad de emoción. [En línea] Disponible en: https://www.saifmohammad.com/. [02 feb. 2020]

Montesinos, G. L. (2014). Análisis de sentimientos y predicción de eventos en twitter.

Montoro, M. A. (2019). Análisis de Sentimientos para la prevención de mensajes de odio en las Redes Sociales.

Mukkamala, R. R., Abid, H., & Vatrapu, R. (Septiembre de 2014). Análisis de sentimientos basado en conjuntos difusos de grandes datos sociales. *En 2014, IEEE 18a Conferencia Internacional de Computación de Objetos Distribuidos Empresariales*, pp. 71-80. IEEE.

Parveen, H., & Pandey, S. (Julio de 2016). Análisis de sentimiento en el conjunto de datos de Twitter usando el algoritmo Naive Bayes. *En 2016, 2ª Conferencia Internacional sobre Computación Aplicada y Teórica y Tecnología de la Comunicación (iCATccT)*, pp. 416-419. IEEE.

Paz, J. P. (26 de Mayo de 2015). Recuperado el 10 de Octubre de 2018, de http://www.jeronimoperez.com/blog/que-es-un-microblog/

Ponce-Cruz, P., Molina, A., and MacCleery, B. (2016). Fuzzy Logic Type 1 and Type 2 Based on LabVIEW™ FPGA. Springer.

Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. Information Processing & Management, 56(6), 102060.

Romsaiyud, W., & Premchaiswadi, W. (2015). Aplicación de técnicas de minería de patrones secuenciales difusos para predecir el liderazgo en las redes sociales. *Conferencia Internacional sobre TIC e Ingeniería del Conocimiento*, pp. 134-137.

RPubs (2017). Evaluación de modelos de Clasificación. Recuperado el 08 de Febrero de 2020, de https://rpubs.com/chzelada/275494

Şahi, H., Kılıç, Y., & Sağlam, R. B. (Septiembre de 2018). Detección automatizada de discursos de odio hacia mujeres en Twitter. *En 2018, 3er Congreso Internacional de Ingeniería y Ciencias de la Computación (UBMK)*, pp. 533-536. IEEE.

Sánchez, E. P. (2017). Análisis de los Procesos de Lematización y Estemizado en Lingüística Computacional.

Setyadi, N., Nasrun, M., & Setianingsih, C. (Diciembre de 2018). Análisis de texto para la detección del habla por odio usando la red neuronal de propagación hacia atrás. *En 2018, Conferencia Internacional sobre Control, Electrónica, Energías Renovables y Comunicaciones (ICCEREC)*, pp. 159-165. IEEE.

Statista. (2022). "Social media y contenido creado por el usuario". [En Línea] https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/. 20 junio 2022.

Taulé, M., & Martí, M. A. (2003). SENSEVAL, una aproximación computacional al significado. *Digithum*, (5).

TALP Research Center. (2021). *Welcome | FreeLing Home Page*. Recuperado el 15 de Agosto de 2020, de Universitat Politécnica de Catalunya: https://nlp.lsi.upc.edu/freeling/node/1

Torrús, A. (18 de Enero de 2018). Jaque a la libertad de expresión. Delitos de odio: el elixir de la nueva Inquisición. *Público*. Recuperado el 29 de Septiembre de 2018, de https://www.publico.es/politica/jaque-libertad-expresion-delitos-odio-elixir-nueva-inquisicion.html

Trim, C. (2018). *El arte de la Tokenización*. Recuperado el 20 de Septiembre de 2019, de IBM : https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en.

Twitter (2019). Developer Platform. Recuperado el 28 de Marzo de 2019, de https://developer.twitter.com/en.

UNAM, G. (01 de Enero de 2018). México, cuarto lugar a nivel mundial en uso de redes sociales. *Excelsior*. Recuperado el 02 de Octubre de 2018, de https://www.excelsior.com.mx/hacker/2018/01/18/1214650

Universal, E. (15 de Septiembre de 2016). Ciberdelitos en México: apuntes sobre este delito que llego para quedarse. *El Universal*. Recuperado el 12 de Noviembre de 2018, de https://www.eluniversal.com.mx/blogs/observatorio-nacional-ciudadano/2016/09/15/ciberdelitos-en-mexico-apuntes-sobre-este-delito

Utami, E., & Luthfi, E. (Marzo de 2018). Minería de texto basada en comentarios de impuestos como análisis de big data usando SVM y selección de características. *En 2018, Conferencia Internacional sobre Tecnologías de la Información y las Comunicaciones (ICOIACT)*, pp. 537-542. IEEE.

Waseem, Z. &. (Junio de 2016). ¿Símbolos odiosos o personas odiosas? Características predictivas para la detección del odio en Twitter. *En Actas del Taller de Investigación para Estudiantes de NAACL*, pp. 88-93.

Watanabe, H. B. (2018). Discurso de odio en Twitter: un enfoque pragmático para recopilar expresiones de odio y ofensivas y realizar la detección de odio del habla. *Acceso IEEE, Vol. 6*, pp. 13825-13835. IEEE.

Zadeh, L.A. (1975). The concept of a linguistic variable and its applications to approximate reasoning. part i, ii, iii. Information Science, 8-9:199–249, 301–357, 43–80.

Zadeh, L. A. (1987). Fuzzy sets and applications. Number 511.32 F8.

Zhang, S., Zhang, C., & Yang, Q. (2003). Preparación de datos para la Minería de Datos. *Inteligencia Artificial Aplicada, Vol.17*(5), pp. 375-381.

Anexo A

A continuación, se muestra el listado de las palabras clave utilizadas para la búsqueda y extracción de comentarios de la red social Twitter.

abandonado	asqueroso	chingue	culpable
aborrecer	ataque	chocar	dañar
abuso	atropello	choque	decrepito
acosar	basura	cobardes	defender
acoso	batalla	colgar	defensa
acusar	bloqueo	combate	defraudar
agresión	bofetada	condenar	delincuente
agresivo	bomba	confinar	demandar
agresor	bombardear	confiscar	denunciar
alboroto	bombardeo	conflicto	depravar
altercado	brusco	confrontar	deprimido
amenazar	brutal	confusión	desafiar
angustia	bruto	conquista	desagradable
aniquilación	burla	conspirador	desalojo
aniquilar	burlón	contra	desastre
ansiedad	cabron	contrabando	descuido
aplastar	callar	contradecir	desgracia
apretar	calumnia	corrupción	desgraciado
arruinar	cáncer	corrupto	deshonesto
asaltante	caos	cortar	deshonrar
asaltar	cárcel	cretino	desigualdad
asalto	castigar	crimen	desistir
asco	castigo	criminal	despreciar
asesinar	catástrofe	criminalidad	desterrar
asesinato	celoso	criticar	destrozar
asesino	censurar	cruel	destruir
asfixia	chantaje	crueldad	detener
detestar	explosivo	infanticidio	mentira
devastador	explotar	inferior	mentiroso
difamación	expulsar	infidelidad	mierda
discordia	falsificación	infiel	miserable

discriminación	feminismo	injusticia	molestar
discriminar	feministas	injusto	mortal
discutir	fraude	inmoral	motín
disgustar	fuerza	insulto	muerte
disparar	furia	inútil	prisión
disputar	furioso	invadir	multitud
distraer	golpear	ira	mutilar
disturbio	gritar	irritable	negar
egoísta	guerra	joder	odio
ejecución	herir	ladrón	ofender
emboscada	hipocresía	látigo	ofendido
empujón	homicida	lesión	ofensa
encarcelar	homicidio	lesionar	paliza
enemigo	homosexual	levantamiento	patear
enfado	horrendo	linchar	pelea
enfurecer	horrible	loco	pendejo
engañar	humillación	maldición	perder
enloquecer	humillar	maldito	perro
enojar	idiota	malo	persecución
escándalo	ilegal	maníaco	perseguir
esclavitud	imbécil	manifestación	perturbar
estallar	incesto	masacre	perverso
estupidez	incitar	masoquismo	pervertir
estúpido	incompetente	matadero	pirata
exagerar	indeciso	matanza	pisotear
excitar	inepto	matar	pistola
explosión	inestable	matón	pobre
prohibir	reproche	tacaño	tumulto
psicópata	revoltoso	temblar	veneno
púdrete	robado	terrorismo	venganza
puñalada	robar	terrorista	vergonzoso
puñetazo	ruptura	tonto	víctima
puto	salvaje	tortura	villano
rabiar	sarcasmo	traición	violar
racista	secuestrar	traicionar	violencia

Anexo B

La lista de Stopwords muestra todas aquellas palabras que son descartadas en el análisis de los comentarios extraídos de la red social Twitter.

a	estado	fuese	hubieran
al	estados	fueseis	hubieras
algo	estamos	fuesen	hubieron
algunas	estando	fueses	hubiese
ante	estar	fui	hubieseis
antes	estará	fuimos	hubiesen
como	estarán	fuiste	hubieses
con	estarás	fuisteis	hubimos
contra	estaré	fuésemos	hubiste
cual	estaréis	ha	hubisteis
cuando	estaría	habida	hubiéramos
de	estaríais	habidas	hubiésemos
del	estarías	habido	hubo
desde	estas	habidos	la
donde	este	habiendo	las
durante	estemos	habremos	le
e	esto	habrá	les
el	estos	habrán	lo
ella	estoy	habrás	los
ellas	estuve	habré	me
ellos	estuviera	habréis	mi
en	estuvierais	habría	mis
entre	estuvieran	habríais	mucho
era	estuvieras	habrían	muchos
erais	estuvieron	habrías	muy
eran	estuviese	habéis	más
eras	estuvieseis	había	mí
tenía	estuviesen	habíais	mía
es	estuvieses	habían	mías
esa	estuviste	habías	yo
esas	estuvisteis	tres	él
ese	estuviésemos	han	éramos
eso	estuvo	has	vuestro
esos	está	hasta	quien
esta	estáis	hay	quién
estaba	están	haya	quienes
estabais	estás	hayamos	quiere
estaban	vuestros	hayan	se
estabas	esté	hayas	sea
estad	estéis	hayáis	sean

estada	estén	he	según
estadas	estés	hemos	tienen
y	fue	hube	sobre
ya	fuera	hubiera	sola
siete	fuerais	hubierais	solamente
mío	serás	teníamos	será
míos	seré	tenían	serán
nada	seréis	tenías	
ni	sería	ti	
no	seríais	tiene	
nos	serías	tienen	
nosotras	seáis	tienes	
nosotros	sido	todo	
nuestra	siendo	todos	
nuestras	sin	tu	
o	sobre	tus	
os	sois	tuve	
otra	somos	tuviera	
otras	son	tuvierais	
otro	soy	tuvieran	
otros	su	tuvieras	
para	sus	tuviese	
pero	suya	tuvieseis	
poco	suyas	tuviesen	
por	suyo	tuvieses	
porque	suyos	tuvisteis	
que	sí	tuviéramos	
quien	también	tuviésemos	
quienes	tanto	tuvo	
qué	te	tuviera	
se	tendrá	tuvierais	
sea	tendrás	tuvieran	
sean	tendré	tuya	
seas	tendréis	tuyas	
seremos	tendría	tuyo	
mío	tendríais	tuyos	
míos	tendrían	tú	
nada	tendrías	un	
ni	tened	una	
no	tenga	uno	
nos	solas	unos	
nuestra	sus	vosotras	
nuestras	tal	teníais	