DETECCIÓN DE PEDERASTIA EN CONVERSACIONES DIGITALES CON PROCESAMIENTO DE LENGUAJE NATURAL

Tesis presentada para obtener el grado de:

Licenciatura en Ingeniería en Ciencias de la Computación

Presenta:

José Luis Jiménez Aguilera

ORCID iD

Correo:

jose.jimenezag@alumno.buap.mx

Director de Tesis y Asesor de Tesis:

Luis Enrique Colmenares Guillen

ORCID iD

Universidad:

Benemérita Universidad Autónoma de Puebla

Facultad:

Ciencias de la Computación

Fecha de Entrega:

Febrero 2024

Agradecimientos

Agradezco a mi persona por el esfuerzo, dedicación y tiempo reflejadas en este documento y proyecto, por esas horas estudiando, analizando e implementando nuevas funcionalidades y realizar un digno trabajo de titulación.

Mi padre por sus consejos, la ayuda financiera y en darme la oportunidad de prepararme profesionalmente en el área que tanto amo.

Mi madre por su cariño, sus ánimos y su apoyo tanto en la cima como en el fondo.

Mi hermano por su compañía, ánimos y el estar presente cuando necesito su ayuda.

Abuela paterna por su calidez, el pensar en cómo ayudarme en todo momento y su apoyo por darme un techo y comida además de mi segunda lengua.

Abuelos maternos por su comprensión, sabiduría, calidez y ser fuente de consejos y conocimiento.

Mis tíos, primos por ser parte esencial y sentirme parte de una familia. familia.

Al laboratorio de software libre por ser mi hogar en todos estos años, el centro de mi aprendizaje y darme la oportunidad de conocer a increíbles personas.

A los compañeros del laboratorio por ser mis amigos en todo mi proceso universitario, siempre los tendré en mi corazón, Enrique, Brisa, Sebastián, Víctor, Juanri, Juanjo, Gonzalo, Gerardo, Charlie, Alberto, Arturo, Ricky, Ángel, Joe

A la maestra Alma Delia Ambrosio por ser mi guía en el área profesional de mi carrera, el ayudarme con mis proyectos, realizar mi servicio social con ella, ayudarme con mi timidez, el hablar en público y ser una gran inspiración y maestra.

Al profesor Luis Enrique Colmenares por su conocimiento, tiempo y apoyo tanto en la realización del articulo científico como en la tesis siempre en camino de la excelencia.

Al profesor Eduardo Ariza por demostrarme que las matemáticas son divertidas y fáciles con los métodos adecuados y demostrarme que puedo con esto y más.

Al profesor Pedro por permitirme ser un lobomentor, el tener el conocimiento de manejar situaciones difíciles y guiar a nuevas generaciones de manera correcta.

A Fernando Ortega, Genesis, Fernando Jiménez, Peter y a todo el equipo de controles por darme la oportunidad de realizar mis prácticas profesionales en General Electric Aviation, enseñarme como se trabaja en la industria y ser mis primeros mentores profesional, siempre estaré agradecido.

A mi pareja, por estar ahí cuando lo necesito, comprenderme, quererme, respetarme y amarme con todo su ser, sé que te gustará este trabajo, vamos a crecer juntos corazón, te amo Pau.

A mis amistades más cercanas por ayudarme a su manera, me llevo todo lo que aprendí con ustedes, superé muchos miedos, estoy aquí por ustedes, no tengo forma de agradecerles, los quiero mucho Mari, Fer, Santhy, Roqueved, Ale, Ken y Candy.

A Gabriela y Octavio por ayudarme con mi salud mental, superar mis obstáculos más difíciles, el ser un ser humano integro y encontrar mi filosofía de vida.

Todos ellos fueron parte vital de mi formación, este trabajo es el más ambicioso que he realizado hasta la fecha, espero que los haga sentir orgullosos.

Just like water, let it flow...

Contenido

Capitulo 1. RESUMEN

En los últimos años donde el estilo de vida está cada vez más contextualizado en un ambiente digital globalizado, con fácil acceso a la información junto sus herramientas y la falta de cultura a las nuevas tecnologías dan como resultado el mal uso y gestión de diferentes medios de comunicación. Un ejemplo de esta problemática es la cual se desarrolla en el presente documento. La pederastia en conversaciones digitales se encuentra en estos casos. El documento muestra el uso de diferentes algoritmos que el Procesamiento del Lenguaje Natural nos puede proporcionar, el proceso para detectar textos de carácter pederasta, la creación de graficas para un entendimiento más visual de la información y el uso de un entorno para el análisis automático y la minería de datos para la medición de la efectividad del proceso; concluyendo en un reporte final sobre el victimario y los resultados del análisis.

Capitulo 2. INTRODUCCIÓN

Con el paso del tiempo, el uso del internet se ha convertido más y más en una herramienta indispensable desde las tareas más complejas hasta las cotidianas, ya sea con dispositivos como celular, PC, laptop, tablet entre otros.

Un ejemplo de lo lejos que ha llegado es la creación desde relojes hasta refrigeradores inteligentes, esto debido a la necesidad y tendencia de un ambiente inteligente, gracias a la posibilidad de automatizar las tareas del hogar, oficina o lugar de trabajo resulta una función muy atractiva, y muchas de estas funciones se personalizan o tienen aún más si se conectan con las redes sociales, potenciando aún más el uso de estas.

Como consecuencia, esto nos lleva a que sectores muy grandes de la población utilicen estas herramientas, en las cuales se encuentran los menores de edad, las cuales utilizan desde navegadores web, servicios de *streaming*, *wikis*, mensajería instantánea y por supuesto las redes sociales.

Una red social es una plataforma digital que conecta a dos o más personas con intereses, relaciones y actividades comunes. Esto le permite conectarse con otras personas e intercambiar ideas. Esta información puede hacerse pública o privada según usted elija [1].

Algunas de las **redes sociales** que se utilizan actualmente son:

- Facebook
- Twitter
- Instagram
- TikTok

La mensajería instantánea, como su nombre indica, es un servicio de comunicación que permite a dos o más personas enviar y recibir mensajes de texto de forma instantánea o instantánea. No hay demora entre la recepción y el envío de mensajes. En el mundo

de la informática internacional, el término mensajería instantánea (IM) se refiere a este servicio [2].

Algunas de las **mensajerías instantáneas** que se utilizan actualmente son:

- WhatsApp
- Discord
- Telegram
- Snapchat
- WeChat
- Facebook Messenger

Las redes sociales y servicios de mensajería instantánea es la principal fuente para distribuir información, fotos, videos, mensajes y mucho más, esto debido a sus sencillas, funciones, cierto anonimato, disponibilidad y su factor instantáneo.

Su uso es variado y puede ser empleado en diferentes áreas, aquí unos ejemplos

- Uso doméstico:
  - Comunicar un mensaje con los familiares
  - Compartir la ubicación de un dispositivo
  - Compartir fotos y videos

- Uso laboral:
  - Compartir noticias
  - Enviar documentos
  - Realizar videollamadas

- Uso General:
  - Realizar *lives*
  - Venta de productos o servicios

Su uso es tal, que las escuelas, grupos, talleres, empresas y muchas más partes de la sociedad utilicen a las redes sociales y servicios de mensajería instantánea como su principal método de comunicación y el lugar para discutir temas varios.

Al ser las redes sociales y servicios de mensajería instantánea relativamente nuevas, no se tiene la cultura ni el conocimiento adecuado para ser utilizadas de manera correcto y mucho menos en un área como Latinoamérica, esto puede llevar a que las actividades delictivas en la web como estafas, software malicioso, difusión de contenido no apto para todo público, piratería, suplantación de identidad, acoso y otras encuentren cómodas, afectando así a la población sin discriminación desde niños hasta adultos mayores.

En México, 50% de las niñas y niños entre 6 y 11 años son usuarios de internet o de una computadora y en el caso de los adolescentes de 12 a 17 años, entre el 80 y 94% usan internet o una computadora [3].

El 25% de los adolescentes entre los 12 y 17 años ha vivido de alguna forma ciberacoso según una encuesta realizada en México. Las autoridades han advertido del aumento de crimines digitales, tráfico de pornografía infantil, entre otros durante los meses de confinamiento [3].

La **violencia en internet** puede suceder cuando:

- Niñas, niños y adolescentes que interactúan con personas con fines sexuales, incluso en redes sociales y plataformas de mensajería.

- Cuando niños son inducidos a realizar actividades como enviar contenidos ofensivos o recibir información perjudicial para su salud.

- El ciberbullying ocurre cuando niños, jóvenes y otras personas difunden rumores, burlas, amenazas y comentarios inapropiados sobre la víctima en las redes sociales.

- Los niños pueden encontrarse en situaciones peligrosas al compartir información personal, contenido multimedia o información familiar [3].

El acoso, comportamiento no deseado y agresivo entre personas en edad escolar que involucran una diferencia entre poder real o percibido. Este se repite o acostumbra a repetirse constantemente. Las dos partes implicadas tanto las personas que son acosadas y las que no pueden padecer problemas duraderos y graves.

Para que se le pueda llamar acoso, el compartimiento debe ser agresivo e incluir:

- **Desequilibrio de poder**: los niños que intimidan usan su poder, como la fuerza física o la popularidad, para controlar o herir a otros. Este desequilibrio de poder puede cambiar con el tiempo y en diferentes situaciones. Incluso si está involucrada la misma persona.

- **Repetitivo**: La conducta de acoso ocurre o tiende a ocurrir varias veces.

El acoso se encuentra en acciones como amenazas, rumores, ataques verbales y físicos, y la exclusión de alguien de un grupo de manera intencional [4].

Internet ha eliminado las barreras socioeconómicas que antes impedían la comunicación, haciendo posible comunicarse en línea con todos, desde amigos y familiares hasta celebridades y líderes mundiales. Los canales abiertos de comunicación suelen ser un elemento positivo para el progreso humano porque fomentan una mayor colaboración y un aprendizaje compartido.

Pero hoy en día, cualquiera que utilice las redes sociales puede sufrir ciberacoso en línea. La naturaleza viral de *Internet* tiene el potencial de cambiar la personalidad y el futuro a largo plazo de una persona en segundos, sin importar quiénes sean o qué experiencias de vida tengan.

El modelo social no se trata de aprender a evitar el ciberacoso, sino de aprender a ser productivo sin permitir que experiencias pasadas, como abusos pasados o posibles abusos, supriman o influyan en los pensamientos y acciones [5].

El acoso por medio de la web no es nuevo, desde los inicios del internet este se encuentra presente. Las maneras que causan gran preocupación son las de carácter sexual, como la pederastia, las redes sociales, que al ser utilizadas por menores de edad y con falta de supervisión adulta son idóneas para estos depredadores, a la falta de cultura o un filtro el cual identifique si la conversación que un menor tenga con otra persona se presente como acoso o no.

Las investigaciones sugieren que 7 de cada 10 jóvenes han sufrido abusos en línea en algún momento de su vida. Si bien el término ciberacoso. suele utilizarse como si fuera un fenómeno independiente, lo cierto es que se trata de una extensión del acoso, un problema que se tiene mucho tiempo conviviendo. El acoso aprovecha trasfondos sociales de prejuicios y discriminación y suele afectar en mayor medida a personas con características protegidas como la raza, la religión, la sexualidad, la identidad de género y la discapacidad [5].

Las niñas, niños y adolescentes están particularmente expuestos a la violencia en internet, la cual puede tener consecuencias graves en su desarrollo, salud mental e integridad personal. Tradicionalmente, el acoso solía limitarse al entorno escolar, y nuestro hogar era concebido como un espacio seguro. Sin embargo, ahora existe la posibilidad de que un joven sufra acoso tanto en el colegio como en el coche familiar o en su casa, estando él solo en su cuarto e incluso ante la presencia de sus padres o tutores y sin que estos adultos se den cuenta.

Puesto que la tecnología de las comunicaciones se encuentra tan sumamente integrada en la vida moderna, los jóvenes tienen pocas posibilidades de escapar de los abusos, y muchos de ellos viven en un estado constante de estrés y ansiedad. Una de cada tres víctimas de acoso se ha autolesionado por este motivo, y 1 de cada 10 ha intentado suicidarse [5].

Capitulo 3. TRABAJOS RELACIONADOS

En esta sección se abarcará todo sobre proyectos, estudios y/o trabajos los cuales son servirán como punto de partida y de comparación, pueden estar relacionados indirecta o directamente a los temas, tecnologías y/o metodología.

Para la detección de pederastas, las herramientas para la identificación de estas prácticas están fuertemente relacionadas, tanto en su enfoque como en procedimientos aplicados con la capacidad de lograr esta detección de forma distinta y propia.

3.1. Proyectos relacionados

3.1.1 NegoBot

Negobot es un agente conversacional que se hace pasar por un niño en chats, redes sociales y otros servicios similares. Como agente conversacional, *Negobot* utiliza técnicas de procesamiento del lenguaje natural (NLP) y recuperación de información (IR), así como inteligencia artificial y aprendizaje automático. Pero la propuesta más innovadora de *Negobot* es tratar la conversación misma como un juego.

La Teoría de Juegos, área de las matemáticas aplicadas que utiliza modelos para estudiar interacciones en estructuras de incentivos formalizadas y llevar a cabo procesos de toma de decisiones. De este modo, *Negobot* propone un juego competitivo en el que se identifican las estrategias óptimas para la consecución del objetivo: obtener información la cual sirva para saber si la persona con la que el agente conversacional está hablando es un pedófilo, y al mismo tiempo analizando las acciones que nos lleven al presunto delincuente a dejar las conversaciones por realizar un comportamiento no sospechoso de no ser un niño real [6].

3.1.2 iCOP

El software, llamado *iCOP*, presentado en la revista *Digital Investigation* de Elsevier combina inteligencia artificial y aprendizaje automático para detectar únicamente contenido ilegal que esté relacionado con menores. De esta forma, quedan excluidos los contenidos pornográficos para adultos. La directora del proyecto, Claudia Peersman, explicó: "La prevalencia de este tipo de redes hace que la detección manual de contenido ilegal sea un desafío casi imposible". Durante el proceso de investigación, se hicieron pruebas sobre casos reales en centros policiales o comisarias. Los resultados mostraron que la herramienta sólo da falsos positivos en un 7,9% y 4,3% en las fotografías localizadas y de vídeos respectivamente. En este sentido, Peersman explicó que el sistema revela quién comparte el contenido y también detecta otros archivos pertenecientes a la misma persona y relacionados con el delito. También afirmó que la herramienta "será muy importante en la lucha contra la pedofilia y el abuso infantil". [7].

3.2. Marco Teórico

Con el trabajo en conjunto de campos como ciencias de la computación, *machine learning* y lenguaje humano, el uso de sus herramientas y algoritmos desarrollados para cada uno resulta en un sin número de usos, los cuales se pueden usar en beneficio de la sociedad actual.

La variedad de herramientas disponibles que se pueden elegir resultara en la metodología y pasos a realizar para resolver un problema, la elección de las siguientes herramientas se realizó con el enfoque de poder procesar texto, trabajar con los datos resultantes, realizar estadísticas y reportes con estos datos.

3.2.1 Procesamiento del Lenguaje Natural (*NLP*)

El Procesamiento del Lenguaje Natural (*NLP* por sus siglas en inglés) es el campo de estudio que se enfoca en la comprensión mediante ordenador del
lenguaje humano. Abarca parte de la Ciencia de Datos, Inteligencia Artificial
(Aprendizaje Automático) y la lingüística.

En *NLP* no es suficiente con comprender meras palabras, se deberá comprender al conjunto de palabras que conforman una oración, y al conjunto de líneas que comprenden un párrafo. Dando un sentido global al
análisis del texto/discurso para poder sacar buenas conclusiones.

El *NLP* tiene varios usos en diferentes áreas, un ejemplo es que en el Resumen de textos, el algoritmo deberá encontrar la idea central de un artículo e ignorar lo que no le sirva, los *ChatBots* deberán ser capaces de mantener una charla con fluidez con el usuario y se deben responder de manera automática, el proceso de Análisis de Sentimientos debe de comprender si un *tweet*, una *review* o algún texto es positivo, negativo y en que magnitud y la clasificación automática de textos en categorías preexistentes o a partir de textos completos, detectar temas que se repiten y crean categorías [8].

3.2.2 Bag of Words

El método conocido como modelo bolsa de palabras (del inglés, *Bag of Words*) utiliza en el procesado del lenguaje ignorando el orden de las palabras para representar documentos, obteniendo una representación de cada documento dependiendo de las palabras encontradas [9].

El modelo *Bag of Words* es muy usado en procesamiento de lenguaje natural y en búsqueda y recuperación de información (IR). También es

usado frecuentemente en métodos de clasificación de documentos donde la frecuencia de aparición de cada palabra es usada como un añadido para entrenar un clasificador.

La frecuencia en *Bag of Words* es una característica común y muy utilizada, es importante el número de apariciones que tiene un término en un texto, no es necesariamente la mejor forma de representar un texto, de igual manera se puede aplicar de manera exitosa en diferentes áreas por ejemplo en el filtrado de correos electrónicos o *emails* [10].

3.2.3 Análisis de sentimientos

El Análisis de sentimiento utiliza procesamiento de lenguaje natural (*NLP*), lingüística computacional y análisis de texto como referencia para poder extraer e identificar información de los recursos que se considere subjetiva.

Es importante resaltar que el tratamiento que se realiza con el Análisis de Sentimiento se basa en relación de asociación y estadística, no se le puede considerar un análisis lingüístico [11].

Al ser el Análisis de sentimiento una subcategoría del procesamiento de lenguaje natural, se le considera parte de las tareas que realiza el procesamiento de lenguaje natural (*NLP*) el cual brinda a las computadoras el poder comprender el lenguaje humano en su formato oral o escrito. Otras tareas del procesamiento de lenguaje natural es reconocer entidades con nombres, resumir texto, identificar idiomas, y muchas otras [12].

3.2.4 Wordcloud

Las *wordcloud* (nubes de palabras) son herramientas útiles para resumir los conceptos más importantes de un texto, una página web o un libro. Cuantas más palabras haya en el texto que se está considerando, más grande aparecerá en la nube de palabras [13].

3.2.5 Python

Python es un lenguaje de programación ampliamente utilizado en aplicaciones web, desarrollo de software, ciencia de datos y aprendizaje automático (ML). Los desarrolladores utilizan *Python* porque es eficiente, fácil de aprender y puede ejecutarse en muchas plataformas diferentes. El software *Python* se puede descargar gratis, se integra bien con cualquier tipo de sistema y acelera el desarrollo [14].

NLTK es la biblioteca más popular para el procesamiento de *NLP*, está escrita en *Python* y tiene una gran comunidad detrás. Una de las ventajas de *NLTK* es su facilidad de uso, de hecho, es la biblioteca más sencilla de utilizar de todas las que existen.

Por otro lado, la plataforma proporciona interfaces para más de 50 recursos léxicos y corpus, como *WordNet*, complementados con varias bibliotecas de procesamiento de texto para codificación, marcado, clasificación, derivación, inferencia semántica y análisis de contenedores para bibliotecas de *NLP* con aplicaciones industriales. y también hay un completo foro de discusión entre los usuarios.

Con tutoriales prácticos que presentan los fundamentos de la programación, así como temas de lingüística computacional, así como una extensa documentación *API*, *NLTK* es ideal para lingüistas, estudiantes, ingenieros, profesores, usuarios industriales en general e investigadores.

Además, la biblioteca está disponible para *Mac OS X*, *Windows* y *Linux*. Además, no debemos olvidar que *NLTK* es un proyecto de código abierto gratuito y, por tanto, impulsado por la comunidad [15].

3.2.6 *Git* y *GitHub*

Sistema de Control de Versiones Distribuido (*DVCS*) que se utiliza para guardar diferentes versiones de un archivo (o conjunto de archivos), con el fin de poder recuperar cualquier versión del archivo que se requiera.

Git también facilita probar y comparar diferentes versiones de un archivo. Esto significa que los detalles sobre qué ha cambiado, quién cambió qué o quién hizo una sugerencia se pueden ver en cualquier momento.

GitHub es una plataforma web donde los usuarios pueden alojar repositorios *Git*. Facilita compartir y colaborar en proyectos con cualquier persona, en cualquier momento [16].

3.2.7 *Web Scrapper*

Durante el *web scraping*, los datos se extraen y almacenan de las páginas web para su análisis o uso en otro lugar. Con este W*eb Scraping* se almacenan diferentes tipos de información: términos de búsqueda, datos de contacto, direcciones de correo electrónico, números de teléfono, o incluso *URL*. Estos pueden almacenarse en una base de datos o tabla local. [17]

Una vez que se ha determinado qué información se requiere y se necesita extraer de qué sitio *web*, se creará un *bot* o *robot*, llamado *Web Scraper*, para extraer datos específicos de una página *web*.

Para que un *robot* se le pueda considerar no malicioso, debe seguir las reglas de salida que se da por el sitio web que desea piratear en el archivo "robots.txt". Para esto se extrae todo el contenido de un sitio web de forma indiscriminada, desde la estructura hasta el contenido.

Esta primera fase se llama *Web Crawling*, y aunque en diversas fuentes se menciona como algo ajeno al *Web Scraping*, en realidad es parte de un proceso mayor. Luego se identifica y extrae el contenido deseado. El último paso es limpiar y formatear los datos.

En este paso, la información extraída se procesa como en el caso del texto y se almacena en archivos de datos estructurados, como *JSON* o *XML*, utilizando un analizador o en objetos *Python*, como *Numpy arrays*, *Pandas DataFrames*, diccionarios, entre otros [18].

Capitulo 4. METODOLOGÍA

La metodología es el conjunto de procedimientos que otorgan las herramientas que se utilizarán para guiar la manera de trabajar, la forma de comunicarse con los miembros del equipo, como presentar los cambios, manejo de tiempo y expectativas en las entregas.

La elección de la metodología depende del tipo de proyecto que se esté utilizando, el número de personas involucradas, las necesidades del proyecto y sus metas.

4.1. Agile

Agile es mucho más que una metodología para el desarrollo de proyectos que precisan de rapidez y flexibilidad es una filosofía que supone una forma distinta de trabajar y de organizarse [19].

La metodología que se seleccionó para este proyecto es la llamada *Scrum*, un proceso para llevar a cabo un conjunto de tareas de forma regular con el objetivo principal de trabajar de manera colaborativa, es decir, para fomentar el trabajo en equipo.

En *Scrum* se van realizando entregas regulares y parciales del trabajo final, de manera prioritaria y en función del beneficio que aportan dichas entregas a los receptores del proyecto. Por este motivo, es una metodología especialmente

indicada para proyectos complejos, con requisitos cambiantes y en los que la innovación y la flexibilidad son protagonistas, ver Figura 1.

1. Planificación: Product Backlog

El *Product Backlog* es la fase en la que se establecen las tareas prioritarias y donde se obtiene información breve y detallada sobre el proyecto que se va a desarrollar.

2. Ejecución: Sprint

Dentro del método *Scrum*, el *Sprint* es el corazón, un intervalo de tiempo que como máximo tiene una duración de un mes y en donde se produce el desarrollo de un producto que es entregable potencialmente. Para entenderlo mejor, si el *Product Owner* solicita el producto se requiere un mínimo esfuerzo para su entrega al cliente.

3. Control: Burn Down

El *Burn Down* es la fase en la que se mide el progreso de un determinado proyecto *Scrum*. En ella, el *Scrum Master* será el encargado de actualizar los gráficos cuando se finalice cada uno de los *Sprint* [20].

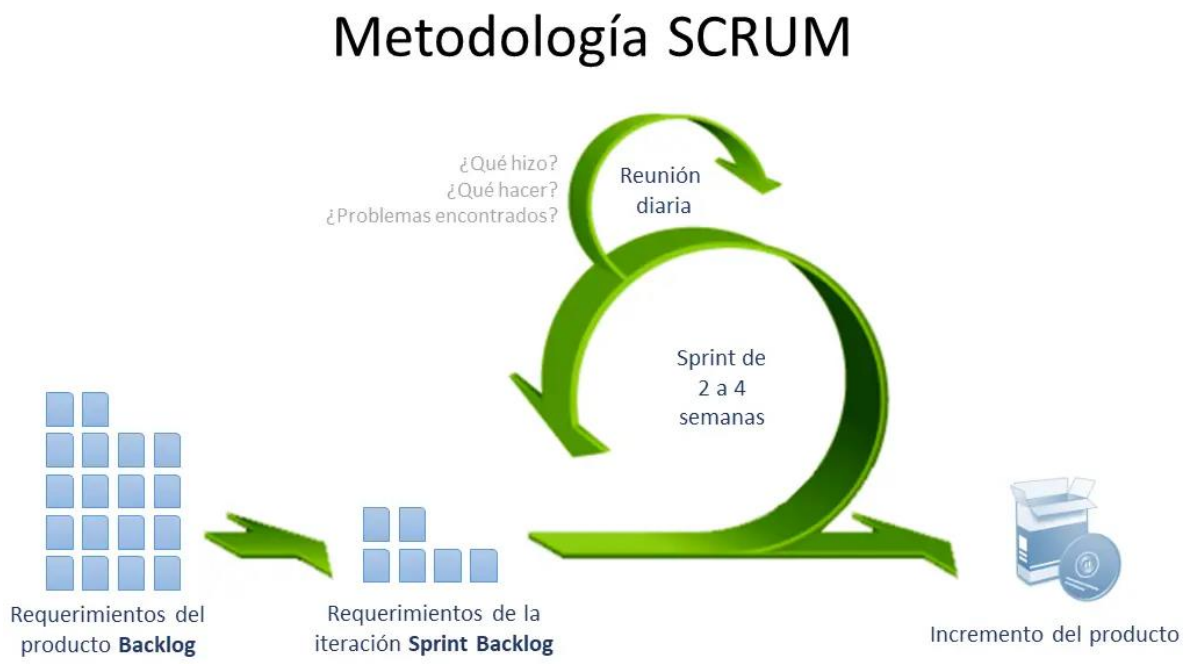

FIGURA 1: METODOLOGÍA SCRUM [21]

Capitulo 5. PROPUESTA

La propuesta es una herramienta capaz de obtener conversaciones de una página web, filtrar los mensajes del agresor, calcular de manera específica datos relevantes como la polaridad de los mensajes, wordclouds los cuales desplegarán de forma visual las palabras que más se utilizan en las conversaciones, documentos de cálculo con el número de apariciones de las palabras utilizadas por pederastas y realizar un reporte final con todos los datos que se obtuvieron.

A continuación, se explica las bases de la aplicación, los pasos que realizan para obtener los resultados, las bibliotecas y tecnologías utilizadas, como se trabajan la integración de todo lo realizado.

5.1. Diagrama de flujo

Modular el programa es algo esencial, el utilizar diferentes tecnologías, métodos, librerías y herramientas se necesita de un documento de apoyo, en este caso se trata del diagrama de flujo, ver Figura 2, ver Figura 3, Figura 4, Figura 5, Figura 6, Figura 7, Figura 8 para ver en detalle el diagrama. En el cual se explica de forma visual el funcionamiento de la herramienta.

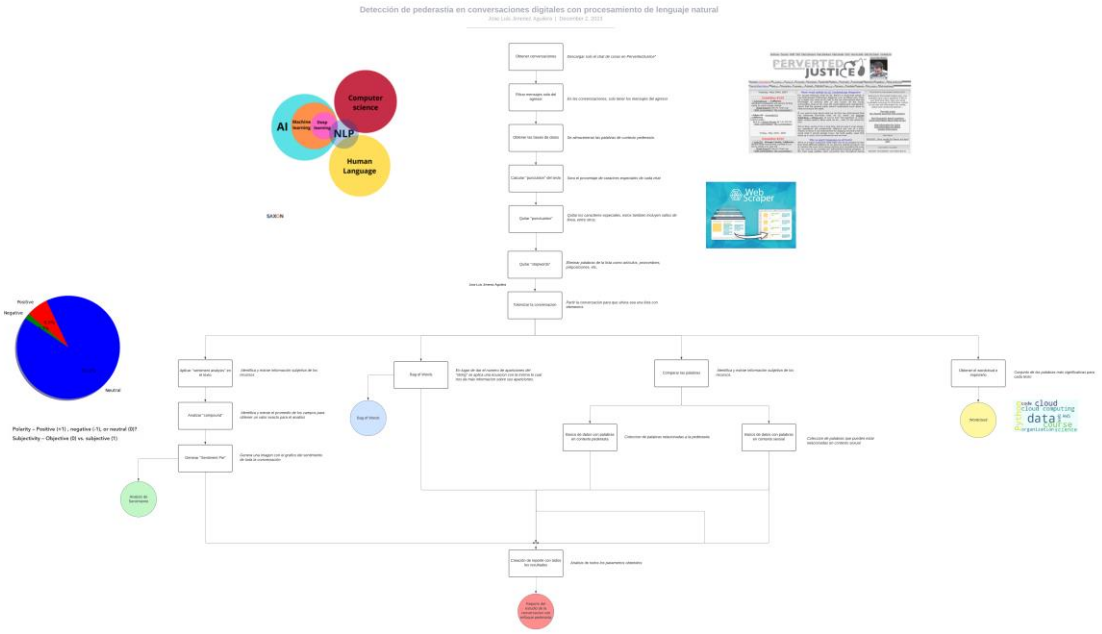

FIGURA 2: DIAGRAMA DE BLOQUE COMPLETO. FUENTE: ELABORACIÓN PROPIA

erastia en conversaciones digitales con procesamiento de lengua

Jose Luis Jimenez Aguilera | December 2, 2023

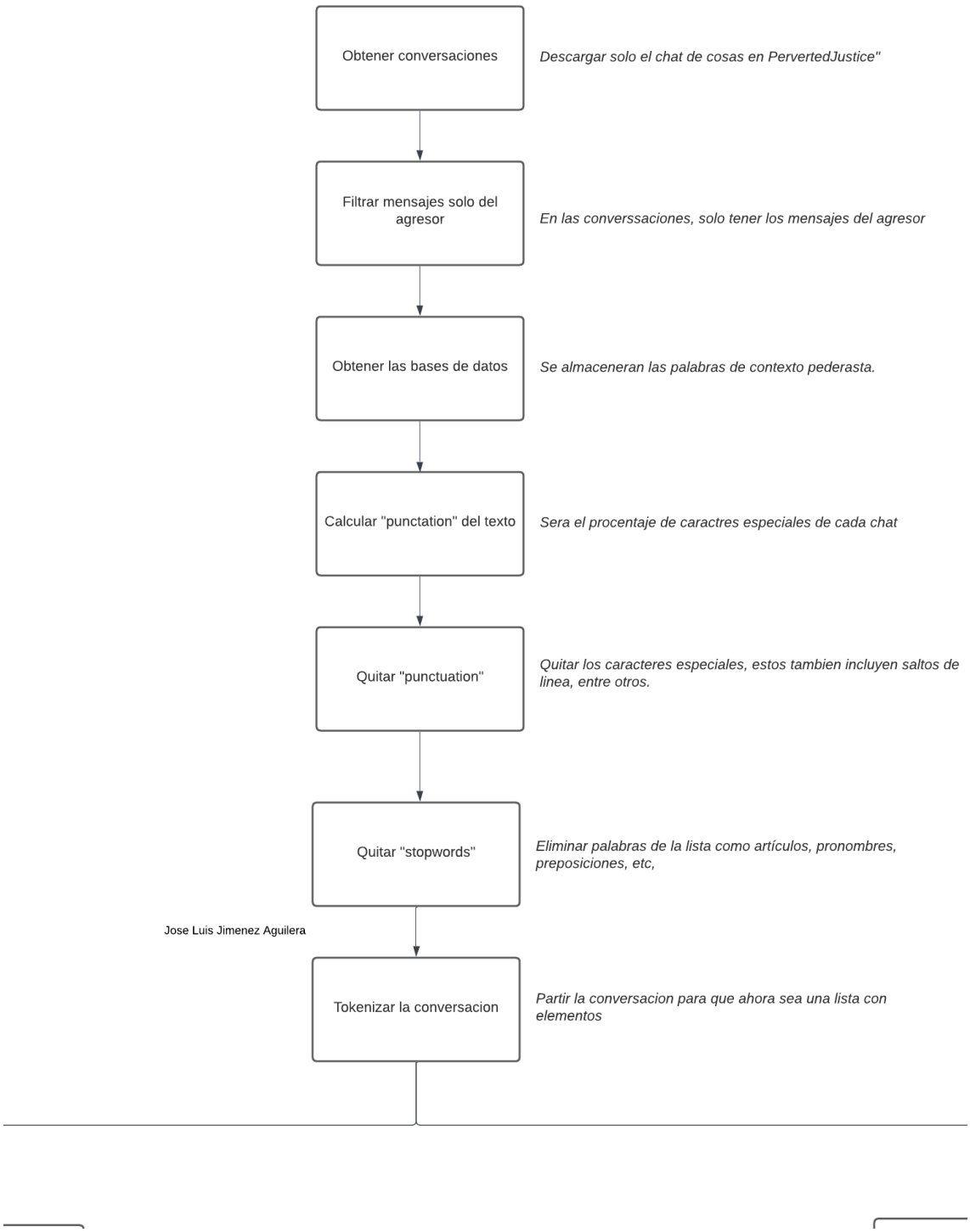

FIGURA 3: DIAGRAMA DE BLOQUE PARTE 1, INICIO. FUENTE: ELABORACIÓN PROPIA

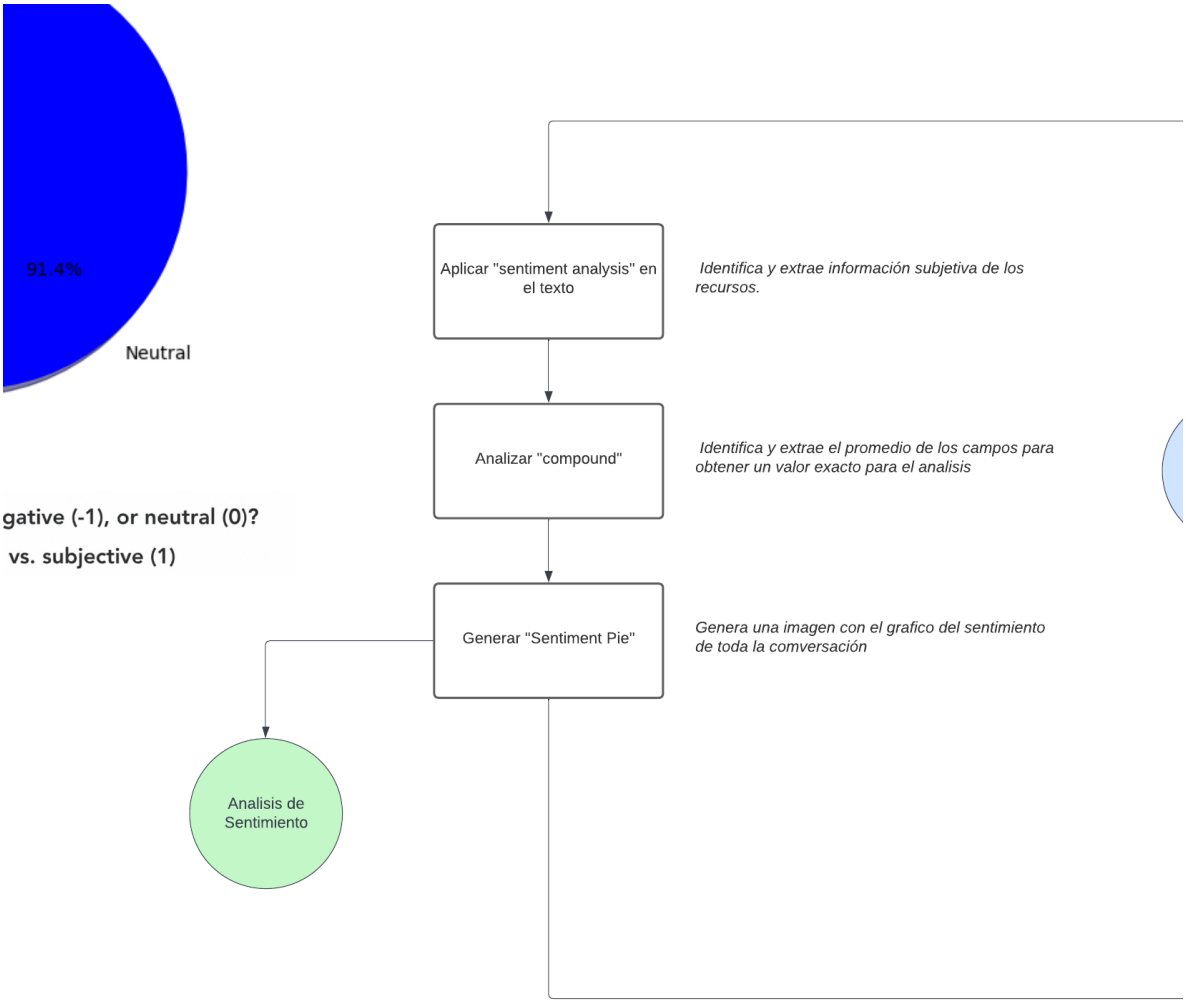

FIGURA 4: DIAGRAMA DE BLOQUE PARTE 2, ANÁLISIS DE SENTIMIENTO. FUENTE: ELABORACIÓN PROPIA

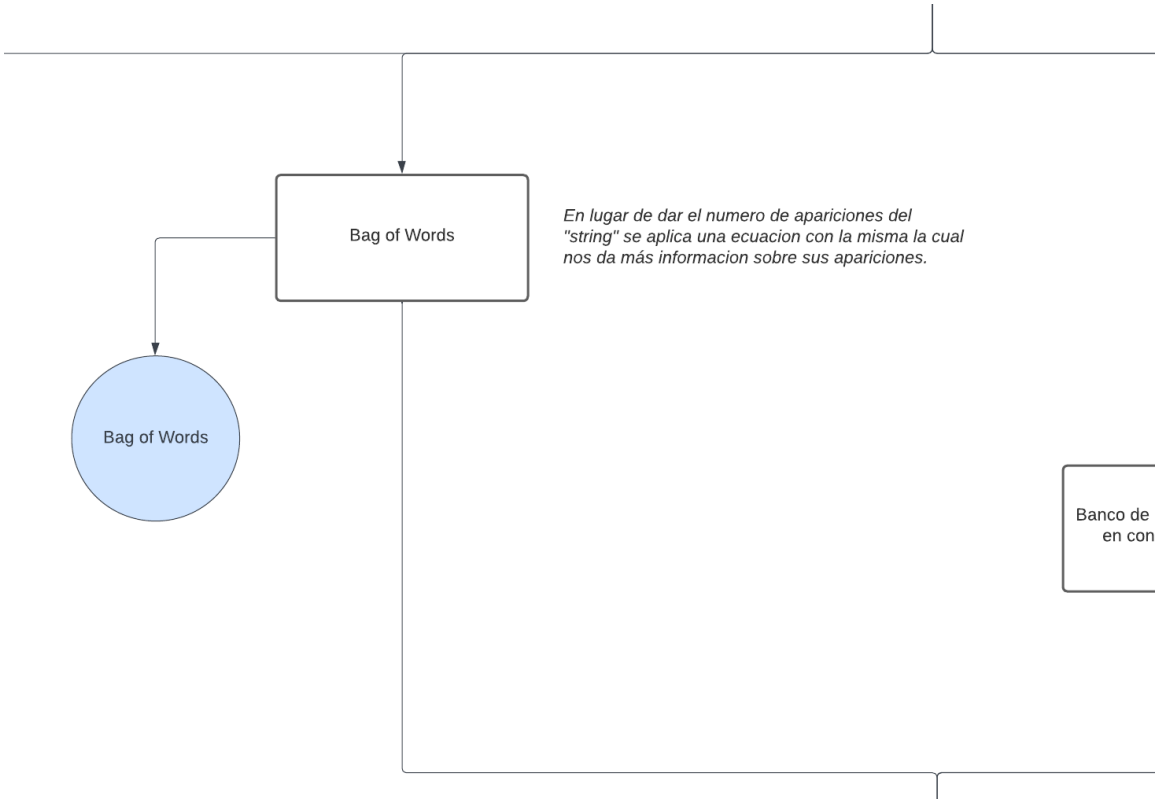

FIGURA 5: DIAGRAMA DE BLOQUE PARTE 3, BAG OF WORDS. FUENTE: ELABORACIÓN PROPIA

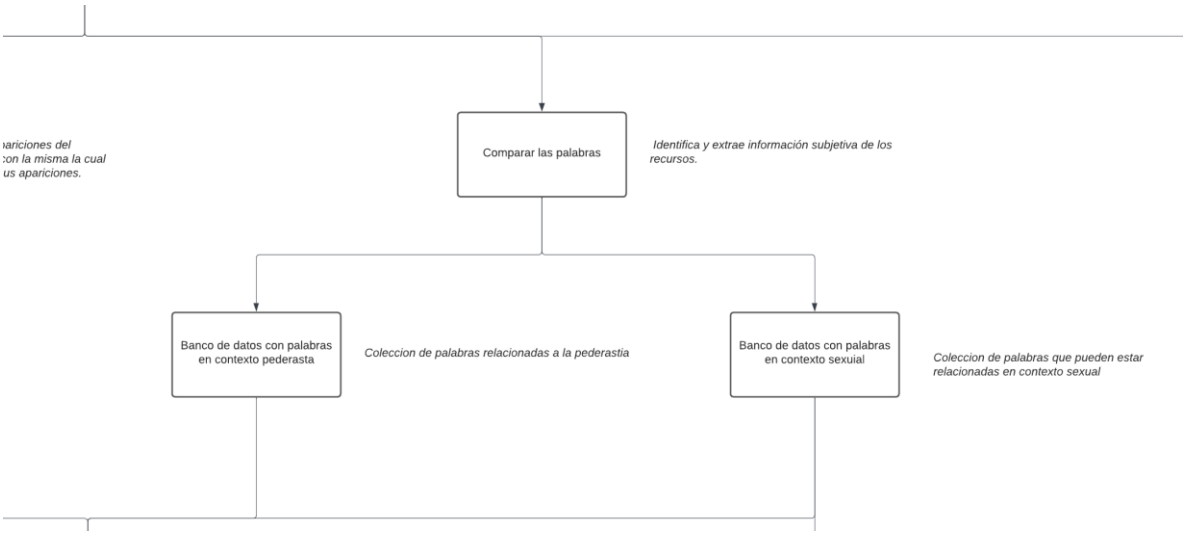

FIGURA 6: DIAGRAMA DE BLOQUE PARTE 4, COMPARACIÓN DE PALABRAS. FUENTE: ELABORACIÓN PROPIA

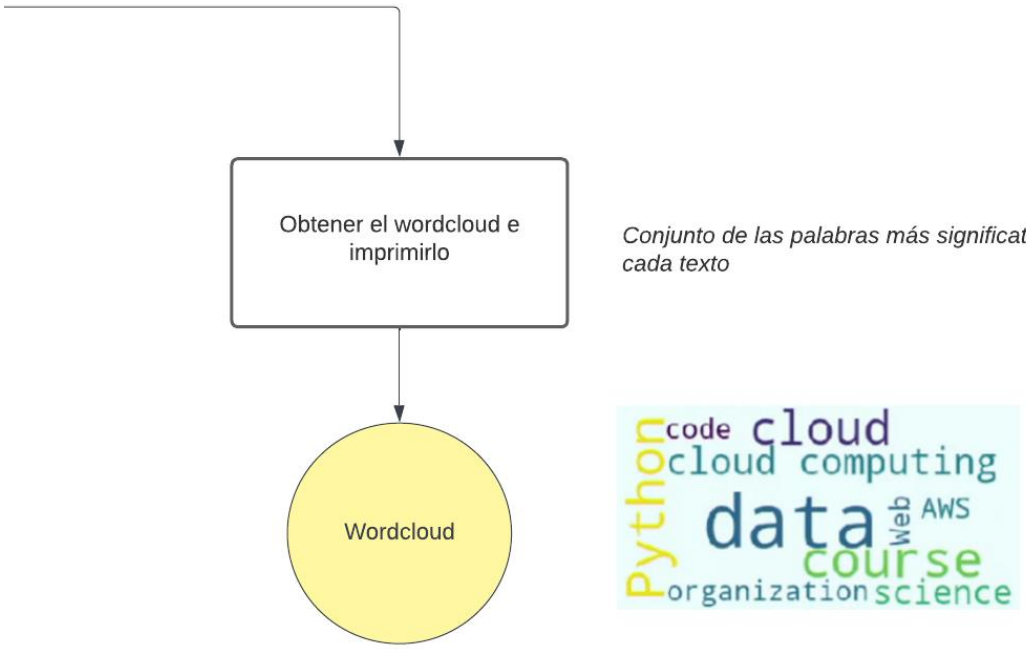

FIGURA 7: DIAGRAMA DE BLOQUE PARTE 5, WORDCLOUD. FUENTE: ELABORACIÓN PROPIA

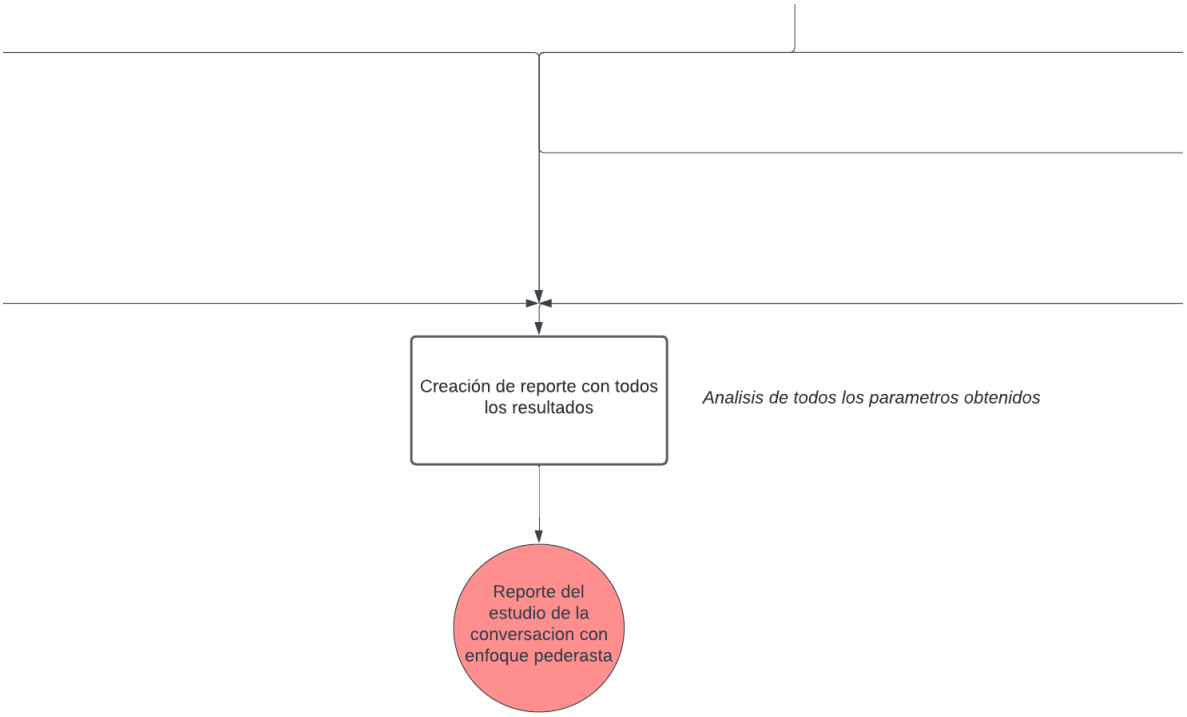

FIGURA 8: DIAGRAMA DE BLOQUE PARTE 6, FINAL. FUENTE: ELABORACIÓN PROPIA

5.2. Planificación de actividades

El programa se divide por segmentos.

5.2.1 Obtención de la conversación por medio de *Perverted Justice*

Proceso que obtiene el código HTML de la página por medio de una técnica llamada *Web Scraping* la cual se utiliza para extraer información de sitios web, teniendo el archivo HTML se filtra los elementos con etiquetas *chatlog* para la conversación y el tercer elemento con la etiqueta *inText* para el nombre el nombre del agresor, ver Figura 9.

FIGURA 9: WEB SCRAPING. FUENTE: ELABORACIÓN PROPIA

5.2.2 Obtención de la base de datos de las palabras con contenido sexual

De dos archivos de texto se obtienen las palabras que se utilizarán, estos archivos contienen palabras de índole pederasta, los archivos se dividen en dos:

- Archivo con palabras usadas por pederastas **sin importar el contexto**
- Archivo con palabras usadas por pederastas **dependiendo el contexto**

Las palabras serán extraídas desde dos archivos *.txt* y almacenadas para su uso más tarde.

5.2.3 Filtrado de los mensajes solo del agresor

Al obtener las conversaciones y el nombre del agresor, se filtra solo los mensajes que sean provenientes del victimario lo cual se realiza al filtrar los mensajes que contengan el respectivo *nickname*, una vez que se obtengan los mensajes se eliminará los datos como *nickname* y la hora, los mensajes de la víctima no tendrán uso y por lo mismo no se almacenarán ni se procesarán, ver Figura 10.

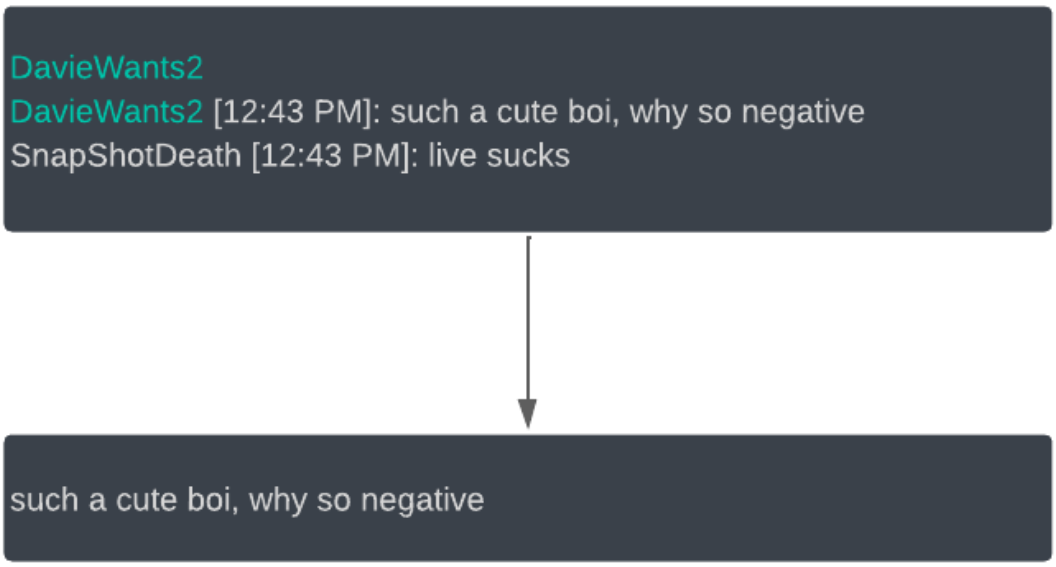

FIGURA 10: FILTRADO DE MENSAJES. FUENTE: ELABORACIÓN PROPIA.

5.2.4 Calcular y remover la *punctation*

La *punctation* son todos los caracteres especiales de un texto, por ejemplo "¿?¡!-_#$%", después de calcular cuánto porcentaje de la conversación se considera *punctation* se procederá a la eliminación de estos caracteres de toda la conversación, esto con el fin de obtener un texto limpio, ver Figura 11.

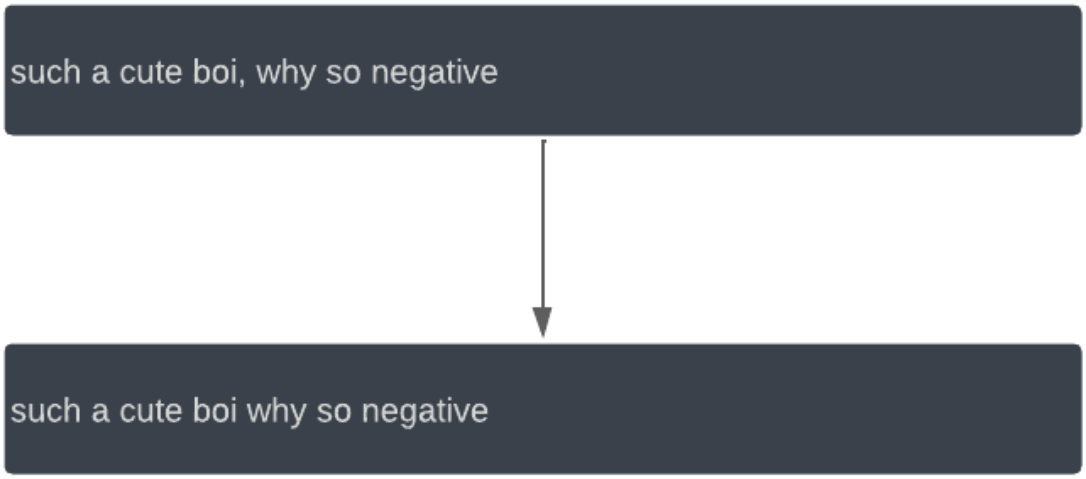

FIGURA 11: ELIMINAR PUNCTATION. FUENTE: ELABORACIÓN PROPIA.

5.2.5 Remover *stop words*

Las *stop words* o palabras vacías son aquellas palabras que se incluyen en el contenido de una página y que no son reconocidas por los robots de *Google*. Algunas palabras vacías como "y" o "de" son muy usadas en el contenido y en las palabras clave, por ello que los motores de búsqueda las eliminan así poder mejorar el rendimiento. [22]
La eliminación de estas palabras es indispensable, así el programa solo procesa las palabras que tienen un sentido de analizar, ver Figura 12.

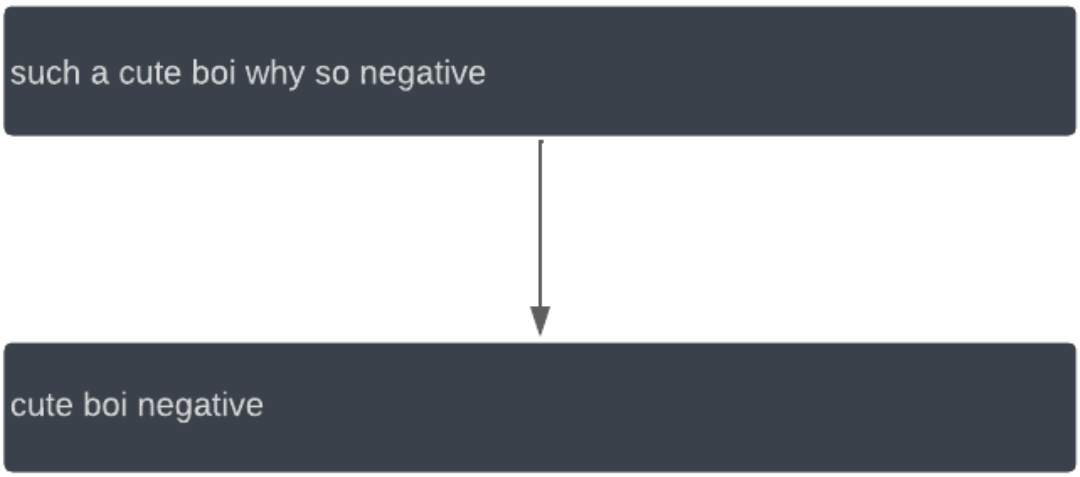

FIGURA 12: ELIMINAR STOPWORDS. FUENTE: ELABORACIÓN PROPIA.

5.2.6 *Tokenizar* la conversación

Nos ayuda a dividir cada mensaje en palabras, así se puede trabajar en cada uno independientemente y realizar los siguientes procesos. Cada palabra se toma como un elemento independiente para su uso en los siguientes pasos, ver Figura 13.

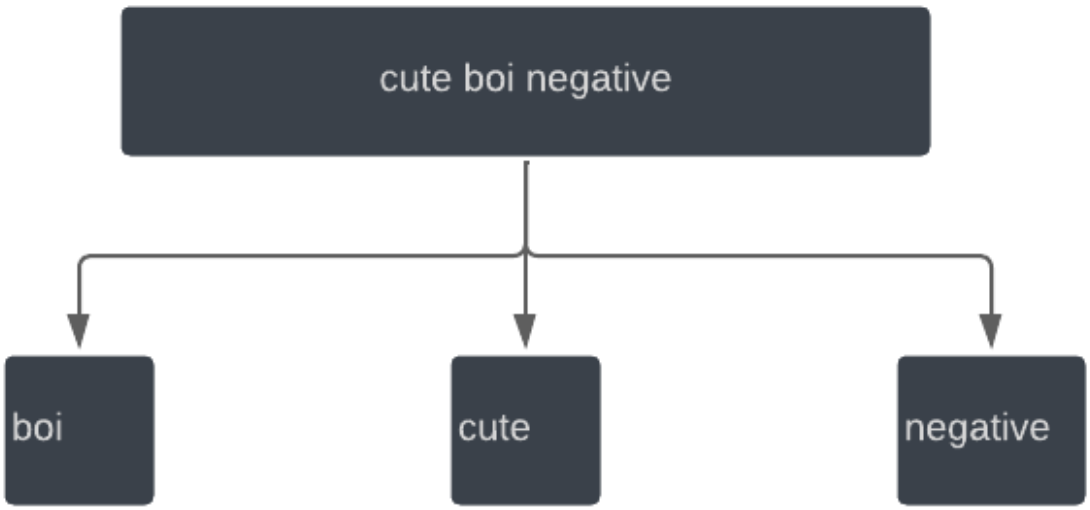

FIGURA 13: TOKENIZAR LA CONVERSACIÓN. FUENTE: ELABORACIÓN PROPIA

5.2.7 Aplicar el análisis de sentimientos

A cada palabra se le aplica el análisis de sentimiento, se obtiene si la palabra tiene un significado neutro, positivo o negativo basándose en el resultado del análisis de sentimiento llamado *Vander*.
Se genera un archivo *xlsx* en el cual se muestra los datos de cada palabra, ver Figura 15.

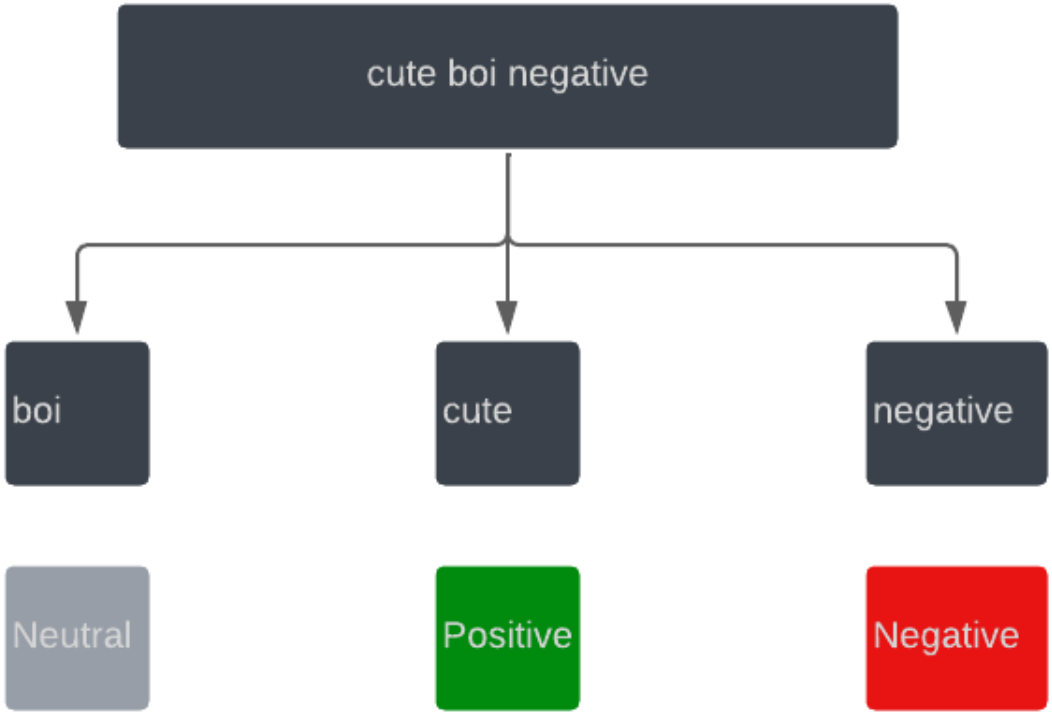

FIGURA 14: ANÁLISIS DE SENTIMIENTO. FUENTE: ELABORACIÓN PROPIA.

5.2.8 Generar *Sentiment Pie*

Con los datos del análisis de sentimientos se obtiene los resultados de todas las palabras de la conversación y se muestra en una gráfica tipo *pie*.
Se genera un archivo *png* con la gráfica y los datos obtenidos, ver Figura 15.

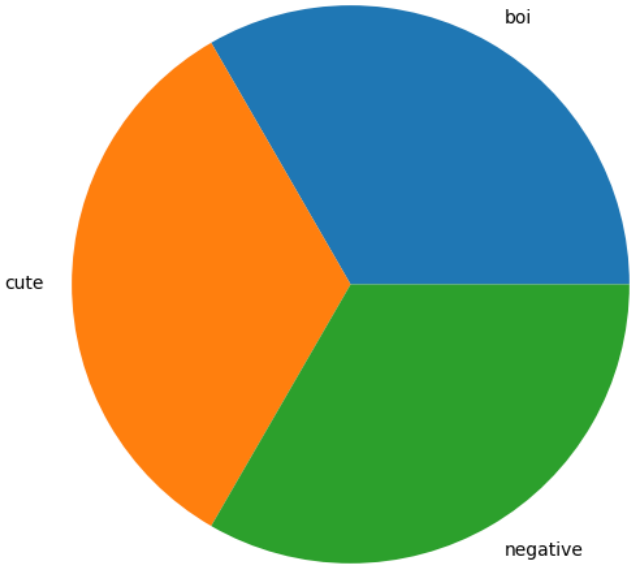

FIGURA 15: SENTIMENT PIE. FUENTE: ELABORACIÓN PROPIA.

5.2.9 Generar el *Bag of Words*

Se genera un recuento de las apariciones de cada palabra en las conversaciones, de igual manera con un archivo de cálculo *xlsx* en el cual se colocan tanto las palabras encontradas como su número de apariciones en el texto, ver Figura 16.

FIGURA 16: BAG OF WORDS. FUENTE: ELABORACIÓN PROPIA.

5.2.10 Comparación de las palabras

Se calcula la similitud entre las palabras en la base de datos relacionadas con pederastia y las palabras encontradas en la conversación.
Se genera un archivo de cálculo *xlsx* con las palabras implicadas, y sus valores, ver Figura 17.

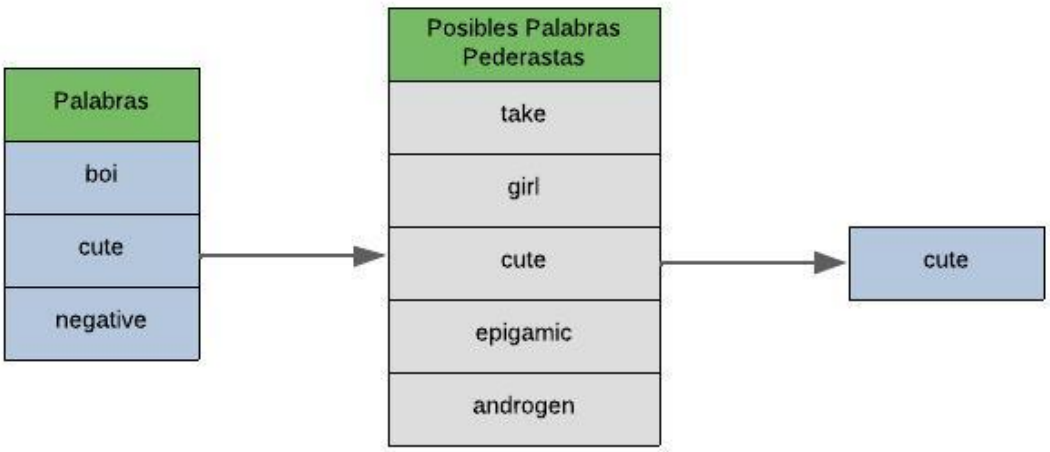

FIGURA 17: COMPARACIÓN DE PALABRAS. FUENTE: ELABORACIÓN PROPIA.

5.2.11 Reporte final

Con los valores de la *punctation*, el análisis de sentimiento, *Bag Of Words* y las comparaciones con la base de datos, se crea un reporte con todos los datos que se obtuvieron de la conversación.

5.3. Base de datos

5.3.1 Obtención de conversaciones pederastas

Perverted Justice (también conocida como "PeeJ") es una organización que se especializa en denunciar a adultos que intentan contactar a menores para tener relaciones sexuales.

El sitio utiliza voluntarios que se hacen pasar por adolescentes de ambos sexos de alrededor de 13 o 14 años, conversan con adultos interesados en intercambiar fotografías en línea e intentan concertar fechas en las que tienen previsto encontrarse.

Los voluntarios son en realidad adultos que recopilan pruebas de comportamiento ilegal para atrapar a estas personas. La información recopilada se publicará en el sitio web y luego podrá ser consultada por cualquier persona [23].

Se utilizó la página *Perverted Justice* como base de datos y se realizó un *script* con *Web Scrapping* para recolectar las conversaciones disponibles en la misma página.

5.3.2 Banco de palabras con contenido pederasta

Colección de palabras con contexto pederasta las cuales serán utilizadas para su comparación con las palabras encontradas en la conversación de *Perverted Justice*.
Estas pueden contener desde propuestas, verbos, adjetivos entre otros.

Existen dos tipos banco de datos el cual el programa utilizará:

- Banco de datos con palabras **utilizadas por pederastas**
- Banco de datos con palabras que **pueden utilizadas por pederastas**

5.4. Tecnologías

Muchas de las funciones del programa necesitan de tecnologías las cuales puedan brindas las funciones y facilidades adecuadas para realizar todas las tareas que el programa requiera.

5.4.1 Python

Python es el lenguaje en el cual se desarrolló en proyecto, debido a la inmensa cantidad de librerías, funciones y soporte, esto facilito mucho el desarrollo ya que muchos de los procesos del programa como filtrado, comparación, creación de archivos, entre otros tienen el soporte suficiente para poder aplicarlos e implementarlos de la forma en la cual el programa y la problemática lo requiere.

5.4.2 Git y GitHub

Al ser un programa extenso, con muchas funciones y archivos, el uso de un controlador de versiones es indispensable, esto ayuda a tener un control de cada versión y cada implementación nueva del equipo, además de tener la opción de compartir el código en caso de ser necesario.

5.4.3 Web Scrapper

Al querer obtener información de una página web de forma automática, el uso de las herramientas como *Web Scrapper* es necesaria, porque se puede obtener el código *HTML* de la página web, en el cual se almacena toda la página, teniendo esto, lo que se realiza es filtrar solo la conversación entre el victimario y la víctima, eliminando los mensajes de la víctima, y así se obtiene finalmente la conversación del victimario.

5.5. Bibliotecas

Las bibliotecas son una colección de segmentos de código las cuales se utilizan para la creación y desarrollo de software, son muy utilizadas debido a que reducen el tiempo de desarrollo de algún programa debido a que las funciones que se necesiten ya están listas para usarse.

Muchas bibliotecas se utilizaron en la aplicación, y son las que se mencionan a continuación.

5.5.1 BeautifulSoup

Beautiful Soup es una biblioteca de *Python* que le permite extraer información de contenido en formato *HTML* o *XML*. Para usarlo, debe especificar un analizador, que es responsable de convertir un documento *HTML* o *XML* en un árbol de objetos *Python* complejo.

Por ejemplo, esto nos permite poder interactuar con los elementos de la página *web* como si estuviéramos usando las herramientas de desarrollo del navegador [24].

5.5.2 nltk

El *Natural Language Toolkit* (*NLTK*) es un conjunto de bibliotecas y programas para *Python* que nos permite realizar muchas tareas relacionadas con el procesamiento del lenguaje natural. La mayoría de las tareas que necesitamos realizar se programarán de manera eficiente en *NLTK* y podremos usarlas directamente en nuestros programas.

Además de los programas, también se distribuyen otros corpus y datos lingüísticos. Esta es una plataforma muy útil tanto para la enseñanza como para el desarrollo y la investigación [25].

5.5.3 Numpy

NumPy es una biblioteca de *Python* especializada en cálculo numérico y análisis de datos, especialmente para grandes volúmenes de datos.

Integra una nueva clase de objetos llamados tablas, que permite representar conjuntos de datos del mismo tipo en múltiples dimensiones y funciones muy eficientes para manipularlos.

La ventaja de *Numpy* sobre las listas predefinidas en *Python* es que procesar matrices es mucho más rápido (hasta 50 veces más rápido) que las listas, lo que lo hace ideal para procesar vectores y matrices grandes [26].

5.5.4 *wordcloud*

La función *WordCloud* en la biblioteca de *wordcloud* permite crear nubes de palabras en *Python*. Esta función tiene varios métodos, la que permite generar una nube de palabras es que se necesitará.

Ten en cuenta que el tamaño de la imagen será 400x200 de forma predeterminada, pero puede personalizar el tamaño con los argumentos de ancho y alto, como en el ejemplo siguiente, o cambiando el tamaño de la imagen con el argumento de escala de forma predeterminada. El valor es 1 y se recomienda para palabras muy largas [27].

5.5.5 *mplib*

Mplib es una biblioteca *Python* liviana que se utiliza para la planificación de movimiento, desarrollada a partir de *ROS* y fácil de preparar. Con unas pocas líneas de comando, puede lograr la mayoría de las funciones de movimiento del robot manipulador [28].

5.5.6 *VaderSentiment*

VaderSentiment es una biblioteca de *Python* que utiliza un enfoque basado en reglas para analizar la opinión en el texto. Este enfoque se basa en un diccionario de palabras y expresiones asociadas con diversos grados de positividad o negatividad.

Utilizando este diccionario, *VaderSentiment* puede asignar un valor numérico que indica el grado de positividad o negatividad expresado en el texto [29].

5.5.7 *matplotlib*

Matplotlib fue desarrollado por John Hunter en 2002. Hunter, neurobiólogo de profesión, creó esta biblioteca con el objetivo de poder observar señales eléctricas en el cerebro de personas con epilepsia. Lo que busco es imitar las funciones gráficas que tiene *Matlab* con *Python*.
Debido a su naturaleza de código abierto, esta biblioteca, a pesar de la muerte de su creador en 2012, continúa logrando avances significativos, hasta el punto de que ahora se puede utilizar para crear gráficos, histogramas, gráficos de barras y muchos otros tipos de gráficos con poco código.
Matplotlib se utiliza en servidores de aplicaciones *web*, *shells* y *scripts* de *Python*, y es especialmente útil para quienes trabajan con *NumPy*.
En esta biblioteca, una forma es una ilustración completa y cada línea de la forma se denomina eje [30].

5.5.8 *hermetrics*

Es una librería que está diseñada para el uso experimental con métricas de palabras. La biblioteca entre otras cosas contiene una clase base la cual es llamada "Métricas", que es altamente configurable y se puede utilizar para la implementación de métricas personalizadas [31].

Los valores comúnmente buscados al utilizar métricas son distancia, distancia normalizada y similitud, que corresponden a los tres métodos principales implementados en *Metric*:

- distancia
- distancia_normalizada
- similitud

Cabe señalar que el objetivo de la *hermétrica* es servir como herramienta de prueba, por lo que el enfoque al implementar métricas que actualmente la incluyen no está en optimizar la complejidad computacional sino en la flexibilidad y la reutilización del código [32].

5.5.9 *os*

El módulo *OS* en *Python* provee funciones para interactuar con el sistema operativo. OS contiene modules de utilidad bajo *Python*. Este módulo provee una forma portable de utilizar funcionalidades dependientes del sistema. Los módulos os y *os.path* incluyen funciones para interactuar junto al sistema [33].

5.5.10 *shutil*

El módulo *shuttil* proporciona una serie de operaciones de alto nivel en archivos y colecciones de archivos. En particular, proporciona funciones que admiten la copia y eliminación de archivos [34].

5.5.11 *pandas*

Pandas es una biblioteca de *Python* especializada en gestionar y analizar estructuras de datos.
Las principales características de esta biblioteca son:

- Define una nueva estructura de datos basada en los *arrays* de la biblioteca *NumPy* pero con nuevas características.
- Le permite leer y escribir fácilmente archivos en bases de datos *CSV*, *Excel* y *SQL*.
- Le permite acceder a datos por índice o nombres de filas y columnas.
- Proporciona métodos para reorganizar, dividir y combinar conjuntos de datos.
- Le permite trabajar con series temporales.
- Realiza todas estas actividades de manera muy eficiente [35]

5.5.12 *pdfkit*

PDFkit es una librería que genera documentos complejos, *multi-pagina* en formato *PDF* de una manera fácil. La *API* incluye funciones de bajo nivel, así como abstracciones para funciones de nivel alto, también genera documentos complejos con unas simples ejecuciones de funciones [36].

5.5.13 *spaCy*

SpaCy, junto con *NLTK*, es una de las bibliotecas más utilizadas en el procesamiento del lenguaje natural (*NLP*). *SpaCy* fue desarrollado por Matt Honnibal y lanzado en 2015; Tiene licencia *MIT* y está disponible en *GitHub*. Esta biblioteca tiene, entre otras cosas, las siguientes características:

- Admite más de 70 idiomas.
- Contiene 80 enlaces traducidos a 24 idiomas.
- Incluye *BERT* previamente capacitado. *BERT* es una arquitectura de aprendizaje profundo basada en transformadores y una de las arquitecturas más potentes disponibles.
- Aprendizaje multitarea.

- Vector de palabras previamente entrenado.
- Fichas de idioma.

Los componentes permiten el reconocimiento de entidades con nombre, etiquetado de partes del discurso, análisis de dependencia, segmentación de oraciones, clasificación de texto, análisis léxico o morfológico, entre otros. Admite modelos personalizados en *PyTorchm Tensorflow* y otros marcos. *SpaCy* incluye modelos previamente entrenados dentro del propio módulo. Incluso se pueden descargar para que la detección de entidades o la extracción de temas (por poner algunos ejemplos) se puedan realizar de forma más automática [37].

5.5.14 Beautifulsoup

Beautiful Soup es una biblioteca de *Python* para analizar documentos *HTML* (incluidos aquellos con marcado incorrecto). Esta biblioteca crea un árbol con todos los elementos del documento y puede usarse para extraer información. Entonces esta biblioteca es muy útil para extraer información de páginas *web* [38].

Capitulo 6. RESULTADOS

En la generación y resultado de todo el procesamiento que la aplicación realiza, se generan diferentes directorios y archivos los cuales reflejan todo el trabajo realizado de tal manera que este ordenado y de fácil lectura, las cuales contienen archivos de formato de cálculo, imágenes, archivo *ARFF* y un reporte final.

6.1. Directorio aggresors

El programa genera diferentes directorios y archivos los cuales reflejan todo el trabajo realizado de tal manera que este ordenado y de fácil lectura.

6.1.1 Directorio "images"

En esta carpeta se guardan todas las imágenes generadas para este agresor que se adjuntaron en el reporte final. "Sentiment Pie", "Wordcloud", "WordCloudPossibleSimilarity" y "WordCloudSimilarity" son los archivos que se encuentran en la carpeta.

6.1.2 Directorio "xlsx"

En esta carpeta se guardan todos los archivos de tipo xlsx que se generaron. "BagOfWords", "BagOfWordsAOS", "PossibleSimilarity" y "Similarity" son los archivos generados.

6.1.3 Archivo "FinalReport.pdf"

Reporte final del análisis que se realizó del agresor, este contiene diferentes secciones las cuales se componen de 1 página explicando la sección y a continuación la sección en particular, las secciones que compone el archivo son las siguientes:

6.1.3.1 Bag of Words

Se genera una tabla con las palabras que se encontraron en la conversación junto con el número de apariciones de las palabras.
La tabla tiene diferentes campos los cuales se describirán a continuación:

- La fila con la etiqueta "Unamed:0" representa las palabras que se están procesando.
- La fila con la etiqueta "0" representa las apariciones de las palabras que se están procesando.

Ver Figura 18 y Figura 19.

	Unnamed: 0	0
0	1	2
1	15	1
2	1880	2
3	1999	1
4	2	1
5	24	1
6	3	1
7	631	2
8	872	2
9	NH	1
10	NY	2
11	REAL	1
12	U	12
13	UP	1
14	YOU	44
15	YOUR	13
16	YOUbecome	1
17	YOu	26
18	abusive	1
19	adore	1
20	adress	1
21	advantage	1
22	ahappy	1
23	almost	1
24	alone	2
25	already	1
26	anal	1
27	andi	1
28	animals	1
29	anyone	2
30	anything	1
31	anyway	1
32	around	4
33	ass	4
34	asses	1
35	away	1
36	back	1
37	bad	1
38	beat	1
39	becareful	1
40	beg	1
41	beside	1
42	bet	4
43	better	1
44	bit	1
45	bitch	1
46	bite	1
47	blodd	1
48	bodies	1

	Unnamed: 0	0
49	body	3
50	boi	80
51	bois	16
52	booted	1
53	bos	1
54	boys	1
55	broke	1
56	bu	2
57	busy	1
58	bye	1
59	came	1
60	cant	3
61	carefully	1
62	casue	2
63	cavity	1
64	chance	1
65	cheep	1
66	cock	9
67	cocks	2
68	complain	1
69	condoms	1
70	control	1
71	corner	2
72	course	1
73	creature	1
74	cry	1
75	cum	2
76	cut	2
77	cute	3
78	dangerous	1
79	david	1
80	day	4
81	demanding	1
82	depressed	1
83	descreeet	1
84	descreet	1
85	didnt	2
86	die	3
87	differenrt	1
88	disbursted	1
89	distributed	1
90	doen	1
91	doesnt	1
92	dont	29
93	drink	1
94	drove	1
95	eagle	1
96	email	1
97	emialed	1

	Unnamed: 0	0
98	enjoy	1
99	even	4
100	ever	2
101	every	1
102	everything	2
103	exactly	1
104	face	3
105	fact	1
106	fag	3
107	fagboi	1
108	faghole	1
109	family	1
110	feel	1
111	feeling	1
112	felt	1
113	feweling	1
114	fill	1
115	filled	1
116	find	3
117	first	1
118	fisrst	1
119	forget	1
120	forgot	1
121	found	1
122	franks	1
123	freak	1
124	fright	1
125	fuck	11
126	fucked	2
127	fucking	1
128	fun	2
129	gaveme	1
130	gentle	1
131	get	11
132	girl	1
133	give	2
134	glad	1
135	go	1
136	good	2
137	got	2
138	grabbed	1
139	great	1
140	green	1
141	guess	2
142	guilt	1
143	guy	1
144	guys	2
145	hand	1

	Unnamed: 0	0
146	handle	1
147	hands	2
148	happens	1
149	happiness	1
150	happy	4
151	hard	3
152	hardly	1
153	hate	1
154	haveing	1
155	head	1
156	hear	3
157	heck	1
158	help	1
159	herd	1
160	hi	1
161	hole	2
162	honda	1
163	hope	4
164	hopeless	1
165	horny	1
166	hours	1
167	huh	1
168	hurt	4
169	id	5
170	ill	6
171	im	6
172	immediately	1
173	infact	1
174	inheritance	1
175	innocent	1
176	inside	1
177	internet	1
178	involved	1
179	invovled	1
180	ish	1
181	islnad	1
182	itd	1
183	ive	1
184	jerk	1
185	k	1
186	keep	1
187	kewl	2
188	kid	11
189	kill	1
190	know	7
191	later	1
192	lesaking	1
193	let	4

	Unnamed: 0	0
194	lets	1
195	lie	1
196	life	4
197	like	26
198	liked	2
199	likes	1
200	line	1
201	little	4
202	live	2
203	lonely	1
204	lonelyness	1
205	long	2
206	look	1
207	looking	1
208	lot	1
209	lots	5
210	lovable	2
211	love	11
212	loved	1
213	made	2
214	make	4
215	makeing	1
216	making	1
217	mancock	1
218	may	3
219	maybe	2
220	mean	1
221	meet	5
222	meeting	1
223	mention	1
224	mentor	1
225	mess	1
226	met	1
227	might	1
228	mind	2
229	minutes	1
230	molest	1
231	mom	1
232	money	1
233	morning	1
234	motel	1
235	mountains	1
236	mountaism	1
237	move	1
238	much	4
239	murder	1
240	must	4
241	mutalate	2
242	nU	1

FIGURA 18: BAG OF WORDS PARTE 1. FUENTE: ELABORACIÓN PROPIA.

	Unnamed: 0	0
243	naked	3
244	nce	1
245	ne	1
246	near	1
247	neck	3
248	need	1
249	needs	1
250	negative	1
251	never	2
252	nice	3
253	nieve	1
254	nite	1
255	nmake	1
256	nope	1
257	nothin	1
258	nothing	1
259	nut	1
260	nuts	4
261	obsesse	1
262	og	1
263	oh	10
264	ok	15
265	okay	1
266	older	1
267	one	5
268	ons	1
269	oops	1
270	open	1
271	opposite	1
272	ot	1
273	parts	1
274	perv	1
275	phone	1
276	piss	1
277	plenty	1
278	poked	1
279	porous	1
280	prefer	1
281	pretty	1
282	pro	1
283	put	2
284	putting	1
285	queens	1
286	rape	3
287	read	1
288	readuy	1
289	ready	1
290	reel	7

	Unnamed: 0	0
291	reely	11
292	relly	1
293	rely	1
294	remind	1
295	resist	2
296	rewward	1
297	ricci	1
298	riccki	1
299	ricki	1
300	ricky	2
301	rickyboi	1
302	rope	1
303	rought	2
304	roughted	1
305	sac	1
306	safe	1
307	save	1
308	say	1
309	scard	1
310	screamer	1
311	screennames	1
312	see	8
313	seeing	1
314	sence	1
315	sex	3
316	show	1
317	shut	1
318	siad	1
319	sickness	1
320	sicko	1
321	silly	2
322	skin	1
323	slice	1
324	slow	1
325	smacked	1
326	smooth	1
327	sorry	6
328	sort	6
329	souther	1
330	spin	1
331	spread	2
332	squermed	1
333	ssaid	1
334	stabbed	1
335	started	1
336	state	2
337	stay	1
338	stick	3

	Unnamed: 0	0
339	still	3
340	stop	1
341	strangel	1
342	stuck	1
343	stupid	1
344	sucked	1
345	suckss	1
346	sure	8
347	surw	1
348	sweet	1
349	take	3
350	taking	1
351	talk	2
352	teach	1
353	tear	1
354	tell	2
355	thankfull	1
356	thanks	1
357	thats	1
358	thepapers	1
359	ther	2
360	thin	1
361	things	2
362	think	6
363	thinking	1
364	thois	1
365	thonk	1
366	three	1
367	thrill	1
368	ti	1
369	tie	1
370	tied	1
371	tif	1
372	time	2
373	times	1
374	timing	1
375	together	1
376	tossed	1
377	town	1
378	tread	1
379	tree	3
380	tried	1
381	trun	1
382	try	1
383	tthe	1
384	turn	2
385	turned	1
386	twisted	1
387	two	1

	Unnamed: 0	0
388	u	5
389	understand	1
390	us	1
391	use	1
392	useless	1
393	vacation	1
394	voice	1
395	vulnarable	1
396	waht	6
397	waiting	1
398	wake	1
399	want	19
400	wanted	6
401	wanting	1
402	wants	2
403	wasa	1
404	watch	1
405	way	1
406	well	12
407	wetting	1
408	wheni	1
409	whose	1
410	wirh	1
411	without	1
412	wont	1
413	work	2
414	world	2
415	worry	3
416	would	1
417	would	7
418	woulds	1
419	wow	2
420	yeah	7
421	yes	4
422	yng	1
423	young	1
424	yup	1

FIGURA 19: BAG OF WORDS PARTE 2. FUENTE: ELABORACIÓN PROPIA.

6.1.3.2 *WordCloud*

Es una forma de representar los datos obtenidos en *Bag Of Words* de forma gráfica, cuanto más presente esté una palabra en el texto considerado, más grande aparecerá en la nube de palabras.
En la Figura 20, se muestra una imagen con un *WordCloud* aplicado en la misma conversación que se utilizó en el *Bag Of Words*.

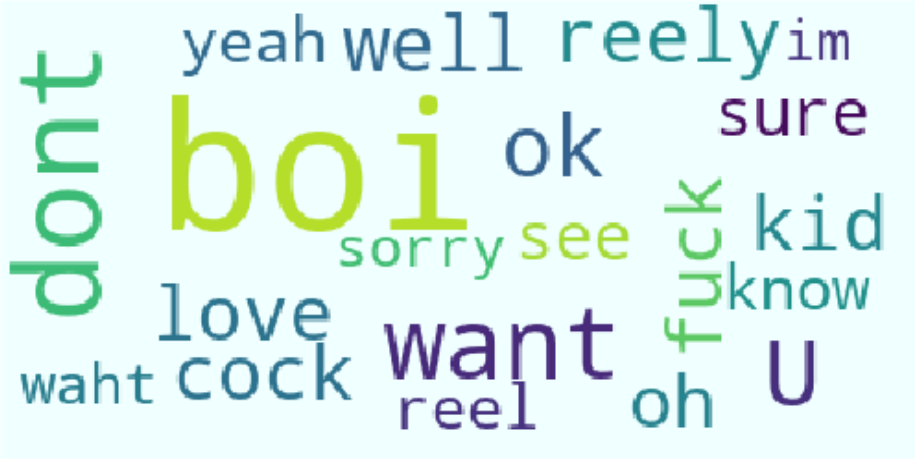

FIGURA 20: WORDCLOUD. FUENTE: ELABORACIÓN PROPIA.

6.1.3.3 *Bag of Words con Análisis de Sentimiento*

Se utilizo el análisis de sentimiento en cada palabra encontrada en la conversación por parte de victimario, para así saber si la palabra se consideró negativa, positiva o neutral.

- La fila con la etiqueta "Unamed:0" representa las palabras que se están procesando.
- La fila con la etiqueta "neg" la cual significa *negative* representa si la palabra es negativa o no, "0" es no negativa y "1" es negativa.
- La fila con la etiqueta "neu" la cual significa *neutral* representa si la palabra es neutral o no, "0" es no neutral y "1" es neutral.
- La fila con la etiqueta "pos" la cual significa *positive* representa si la palabra es positiva o no, "0" es no positiva y "1" es positiva.
- La fila con la etiqueta "compound" la cual significa el valor en promedio de las palabras:

- Si el valor es **mayor que 0.5** significa que es mayormente **positivo**.
- Si es **menor que 0.5** significa que es mayormente **neutral**.
- Si es **menor igual que -0.5** significa que es mayormente **negativo**.

Ver Figura 21, Figura 22 y Figura 23.

	Unnamed: 0	neg	neu	pos	compound
0	cute	0	0	1	0.4588
1	boi	0	1	0	0.0000
2	negative	1	0	0	-0.5719
3	lots	0	1	0	0.0000
4	things	0	1	0	0.0000
5	kid	0	1	0	0.0000
6	dont	0	1	0	0.0000
7	let	0	1	0	0.0000
8	get	0	1	0	0.0000
9	YOU	0	1	0	0.0000
10	sorry	1	0	0	-0.0772
11	life	0	1	0	0.0000
12	suckss	0	1	0	0.0000
13	reely	0	1	0	0.0000
14	make	0	1	0	0.0000
15	YOu	0	1	0	0.0000
16	reel	0	1	0	0.0000
17	happy	0	0	1	0.5719
18	taking	0	1	0	0.0000
19	makeing	0	1	0	0.0000
20	little	0	1	0	0.0000
21	fagboi	0	1	0	0.0000
22	adore	0	0	1	0.5574
23	needs	0	1	0	0.0000
24	yeah	0	0	1	0.2960
25	plenty	0	1	0	0.0000
26	inheritance	0	1	0	0.0000
27	yes	0	0	1	0.4019
28	course	0	1	0	0.0000
29	see	0	1	0	0.0000
30	bois	0	1	0	0.0000
31	naked	0	1	0	0.0000
32	bodies	0	1	0	0.0000
33	waht	0	1	0	0.0000
34	turn	0	1	0	0.0000
35	hard	1	0	0	-0.1027
36	want	0	0	1	0.0772
37	well	0	0	1	0.2732
38	ons	0	1	0	0.0000
39	like	0	0	1	0.3612
40	watch	0	1	0	0.0000
41	im	0	1	0	0.0000
42	looking	0	1	0	0.0000
43	asses	0	1	0	0.0000
44	wow	0	0	1	0.5859
45	relly	0	1	0	0.0000
46	long	0	1	0	0.0000
47	ishad	0	1	0	0.0000
48	young	0	1	0	0.0000

	Unnamed: 0	neg	neu	pos	compound
49	wetting	0	1	0	0.0000
50	meeting	0	1	0	0.0000
51	rickyboi	0	1	0	0.0000
52	timing	0	1	0	0.0000
53	everything	0	1	0	0.0000
54	casue	0	1	0	0.0000
55	wanted	0	1	0	0.0000
56	fisrst	0	1	0	0.0000
57	ricky	0	1	0	0.0000
58	already	0	1	0	0.0000
59	waiting	0	1	0	0.0000
60	doesnt	0	1	0	0.0000
61	even	0	1	0	0.0000
62	know	0	1	0	0.0000
63	horny	0	1	0	0.0000
64	gaverne	0	1	0	0.0000
65	adress	0	1	0	0.0000
66	15	0	1	0	0.0000
67	bet	0	1	0	0.0000
68	smooth	0	1	0	0.0000
69	fag	1	0	0	-0.4767
70	ass	1	0	0	-0.5423
71	sure	0	0	1	0.3182
72	would	0	1	0	0.0000
73	love	0	0	1	0.6369
74	tear	0	1	0	0.0000
75	UP	0	1	0	0.0000
76	U	0	1	0	0.0000
77	meet	0	1	0	0.0000
78	line	0	1	0	0.0000
79	older	0	1	0	0.0000
80	guys	0	1	0	0.0000
81	wont	0	1	0	0.0000
82	hurt	1	0	0	-0.5267
83	kewl	0	0	1	0.3182
84	still	0	1	0	0.0000
85	ill	1	0	0	-0.4215
86	back	0	1	0	0.0000
87	ricki	0	1	0	0.0000
88	wants	0	1	0	0.0000
89	2	0	1	0	0.0000
90	guess	0	1	0	0.0000
91	thonk	0	1	0	0.0000
92	money	0	1	0	0.0000
93	rewward	0	1	0	0.0000
94	YOUbecome	0	1	0	0.0000
95	sex	0	1	0	0.0000
96	must	0	1	0	0.0000
97	ne	0	1	0	0.0000

	Unnamed: 0	neg	neu	pos	compound
98	descreeet	0	1	0	0.0000
99	queens	0	1	0	0.0000
100	souther	0	1	0	0.0000
101	state	0	1	0	0.0000
102	spin	0	1	0	0.0000
103	ricki	0	1	0	0.0000
104	show	0	1	0	0.0000
105	work	0	1	0	0.0000
106	alone	1	0	0	-0.2500
107	ok	0	0	1	0.2960
108	cant	0	1	0	0.0000
109	resist	0	1	0	0.0000
110	phone	0	1	0	0.0000
111	fucked	1	0	0	-0.6597
112	rape	1	0	0	-0.6908
113	REAL	0	1	0	0.0000
114	tell	0	1	0	0.0000
115	put	0	1	0	0.0000
116	rope	0	1	0	0.0000
117	around	0	1	0	0.0000
118	YOUR	0	1	0	0.0000
119	neck	0	1	0	0.0000
120	fuck	1	0	0	-0.5423
121	sweet	0	0	1	0.4588
122	hear	0	1	0	0.0000
123	beg	0	1	0	0.0000
124	die	1	0	0	-0.5994
125	worry	1	0	0	-0.4404
126	mind	0	1	0	0.0000
127	rought	0	1	0	0.0000
128	NY	0	1	0	0.0000
129	live	0	1	0	0.0000
130	exactly	0	1	0	0.0000
131	3	0	1	0	0.0000
132	ish	0	1	0	0.0000
133	green	0	1	0	0.0000
134	1999	0	1	0	0.0000
135	honda	0	1	0	0.0000
136	corner	0	1	0	0.0000
137	sort	0	1	0	0.0000
138	freak	1	0	0	-0.4404
139	yng	0	1	0	0.0000
140	might	0	1	0	0.0000
141	one	0	1	0	0.0000
142	great	0	0	1	0.6249
143	anal	0	1	0	0.0000
144	take	0	1	0	0.0000
145	chance	0	0	1	0.2500

FIGURA 21: BAG OF WORDS CON ANALISIS DE SENTIMIENTO PARTE 1. FUENTE: ELABORACIÓN PROPIA .

	Unnamed: 0	neg	neu	pos	compound
146	1	0	1	0	0.0000
147	631	0	1	0	0.0000
148	872	0	1	0	0.0000
149	1880	0	1	0	0.0000
150	david	0	1	0	0.0000
151	franks	0	1	0	0.0000
152	feel	0	1	0	0.0000
153	ahappy	0	1	0	0.0000
154	anything	0	1	0	0.0000
155	stupid	1	0	0	-0.5267
156	lie	0	1	0	0.0000
157	siad	0	1	0	0.0000
158	ther	0	1	0	0.0000
159	sicko	0	1	0	0.0000
160	becareful	0	1	0	0.0000
161	tread	0	1	0	0.0000
162	carefully	0	0	1	0.1280
163	itd	0	1	0	0.0000
164	dangerous	1	0	0	-0.4767
165	three	0	1	0	0.0000
166	nothin	0	1	0	0.0000
167	never	0	1	0	0.0000
168	oh	0	1	0	0.0000
169	infact	0	1	0	0.0000
170	opposite	0	1	0	0.0000
171	give	0	1	0	0.0000
172	think	0	1	0	0.0000
173	rely	0	1	0	0.0000
174	nmake	0	1	0	0.0000
175	id	0	1	0	0.0000
176	ever	0	1	0	0.0000
177	got	0	1	0	0.0000
178	better	0	0	1	0.4404
179	email	0	1	0	0.0000
180	say	0	1	0	0.0000
181	liked	0	0	1	0.4215
182	almost	0	1	0	0.0000
183	immediately	0	1	0	0.0000
184	perv	0	1	0	0.0000
185	find	0	1	0	0.0000
186	vulnarable	0	1	0	0.0000
187	thankfull	0	1	0	0.0000
188	hope	0	0	1	0.4404
189	happiness	0	0	1	0.5574
190	happens	0	1	0	0.0000
191	world	0	1	0	0.0000
192	safe	0	0	1	0.4404
193	innocent	0	0	1	0.3400

	Unnamed: 0	neg	neu	pos	compound
194	wake	0	1	0	0.0000
195	read	0	1	0	0.0000
196	thepapers	0	1	0	0.0000
197	internet	0	1	0	0.0000
198	porous	0	1	0	0.0000
199	nice	0	0	1	0.4215
200	ive	0	1	0	0.0000
201	time	0	1	0	0.0000
202	hate	1	0	0	-0.5719
203	obsesse	0	1	0	0.0000
204	nce	0	1	0	0.0000
205	girl	0	1	0	0.0000
206	forget	1	0	0	-0.2263
207	herd	0	1	0	0.0000
208	try	0	1	0	0.0000
209	gentle	0	0	1	0.4404
210	lovable	0	0	1	0.6124
211	help	0	0	1	0.4019
212	depressed	1	0	0	-0.5106
213	silly	0	0	1	0.0258
214	enjoy	0	0	1	0.4939
215	seeing	0	1	0	0.0000
216	fright	1	0	0	-0.3818
217	face	0	1	0	0.0000
218	putting	0	1	0	0.0000
219	hands	0	1	0	0.0000
220	thin	0	1	0	0.0000
221	molest	1	0	0	-0.4767
222	guy	0	1	0	0.0000
223	huh	0	1	0	0.0000
224	making	0	1	0	0.0000
225	sence	0	1	0	0.0000
226	two	0	1	0	0.0000
227	minutes	0	1	0	0.0000
228	didnt	0	1	0	0.0000
229	advantage	0	0	1	0.2500
230	nieve	0	1	0	0.0000
231	differenrt	0	1	0	0.0000
232	maybe	0	1	0	0.0000
233	voice	0	1	0	0.0000
234	may	0	1	0	0.0000
235	body	0	1	0	0.0000
236	talk	0	1	0	0.0000
237	times	0	1	0	0.0000
238	hand	0	0	1	0.4939
239	fact	0	1	0	0.0000
240	tried	0	1	0	0.0000
241	bite	0	1	0	0.0000
242	cock	1	0	0	-0.1531

	Unnamed: 0	neg	neu	pos	compound
243	away	0	1	0	0.0000
244	strangel	0	1	0	0.0000
245	wasa	0	1	0	0.0000
246	cheep	0	1	0	0.0000
247	thrill	0	0	1	0.3612
248	felt	0	1	0	0.0000
249	bad	1	0	0	-0.5423
250	lonely	1	0	0	-0.3612
251	mean	0	1	0	0.0000
252	look	0	1	0	0.0000
253	u	0	1	0	0.0000
254	kill	1	0	0	-0.6908
255	screamer	0	1	0	0.0000
256	much	0	1	0	0.0000
257	thats	0	1	0	0.0000
258	sickness	0	1	0	0.0000
259	started	0	1	0	0.0000
260	first	0	1	0	0.0000
261	mutalate	0	1	0	0.0000
262	animals	0	1	0	0.0000
263	og	0	1	0	0.0000
264	guilt	1	0	0	-0.2732
265	lonelyness	0	1	0	0.0000
266	hopeless	1	0	0	-0.4588
267	haveing	0	1	0	0.0000
268	hole	0	1	0	0.0000
269	fill	0	1	0	0.0000
270	bos	0	1	0	0.0000
271	faghole	0	1	0	0.0000
272	cum	0	1	0	0.0000
273	day	0	1	0	0.0000
274	tie	0	1	0	0.0000
275	24	0	1	0	0.0000
276	hours	0	1	0	0.0000
277	need	0	1	0	0.0000
278	lesaking	0	1	0	0.0000
279	motel	0	1	0	0.0000
280	descreet	0	1	0	0.0000
281	mention	0	1	0	0.0000
282	anyone	0	1	0	0.0000
283	oops	0	1	0	0.0000
284	booted	0	1	0	0.0000
285	teach	0	1	0	0.0000
286	mentor	0	1	0	0.0000
287	jerk	1	0	0	-0.3400
288	move	0	1	0	0.0000
289	fucking	0	1	0	0.0000
290	lets	0	1	0	0.0000

FIGURA 22: BAG OF WORDS CON ANALISIS DE SENTIMIENTO PARTE 2. FUENTE: ELABORACIÓN PROPIA .

	Unnamed: 0	neg	neu	pos	compound
291	slow	0	1	0	0.0000
292	prefer	0	1	0	0.0000
293	control	0	1	0	0.0000
294	without	0	1	0	0.0000
295	condoms	0	1	0	0.0000
296	demanding	1	0	0	-0.2263
297	andi	0	1	0	0.0000
298	complain	1	0	0	-0.3612
299	drink	0	1	0	0.0000
300	piss	1	0	0	-0.4019
301	morning	0	1	0	0.0000
302	bitch	1	0	0	-0.5859
303	hi	0	1	0	0.0000
304	understand	0	1	0	0.0000
305	morn	0	1	0	0.0000
306	nU	0	1	0	0.0000
307	yup	0	1	0	0.0000
308	ricci	0	1	0	0.0000
309	surw	0	1	0	0.0000
310	doen	0	1	0	0.0000
311	wirh	0	1	0	0.0000
312	spread	0	1	0	0.0000
313	readuy	0	1	0	0.0000
314	mancock	0	1	0	0.0000
315	filled	0	1	0	0.0000
316	scard	0	1	0	0.0000
317	bu	0	1	0	0.0000
318	tif	0	1	0	0.0000
319	woulds	0	1	0	0.0000
320	loved	0	0	1	0.5994
321	near	0	1	0	0.0000
322	forgot	0	1	0	0.0000
323	ti	0	1	0	0.0000
324	save	0	0	1	0.4939
325	nuts	1	0	0	-0.3182
326	keep	0	1	0	0.0000
327	involved	0	1	0	0.0000
328	handle	0	1	0	0.0000
329	beat	0	1	0	0.0000
330	remind	0	1	0	0.0000
331	ot	0	1	0	0.0000
332	pro	0	1	0	0.0000
333	invovled	0	1	0	0.0000
334	murder	1	0	0	-0.6908
335	together	0	1	0	0.0000
336	us	0	1	0	0.0000
337	family	0	1	0	0.0000
338	fun	0	0	1	0.5106

	Unnamed: 0	neg	neu	pos	compound
339	stick	0	1	0	0.0000
340	cocks	0	1	0	0.0000
341	cut	1	0	0	-0.2732
342	use	0	1	0	0.0000
343	feweling	0	1	0	0.0000
344	blodd	0	1	0	0.0000
345	feeling	0	0	1	0.1280
346	inside	0	1	0	0.0000
347	boys	0	1	0	0.0000
348	cavity	0	1	0	0.0000
349	open	0	1	0	0.0000
350	lot	0	1	0	0.0000
351	wheni	0	1	0	0.0000
352	k	0	1	0	0.0000
353	bye	0	1	0	0.0000
354	busy	0	1	0	0.0000
355	town	0	1	0	0.0000
356	shut	0	1	0	0.0000
357	screennames	0	1	0	0.0000
358	mountains	0	1	0	0.0000
359	NH	0	0	1	0.4939
360	nope	0	1	0	0.0000
361	mountaism	0	1	0	0.0000
362	thinking	0	1	0	0.0000
363	met	0	1	0	0.0000
364	abusive	1	0	0	-0.6369
365	likes	0	0	1	0.4215
366	vacation	0	1	0	0.0000
367	nothing	0	1	0	0.0000
368	glad	0	0	1	0.4588
369	emialed	0	1	0	0.0000
370	heck	0	1	0	0.0000
371	useless	1	0	0	-0.4215
372	creature	0	1	0	0.0000
373	anyway	0	1	0	0.0000
374	mess	1	0	0	-0.3612
375	squermed	0	1	0	0.0000
376	good	0	0	1	0.4404
377	tthe	0	1	0	0.0000
378	whose	0	1	0	0.0000
379	grabbed	0	1	0	0.0000
380	twisted	0	1	0	0.0000
381	tossed	0	1	0	0.0000
382	thois	0	1	0	0.0000
383	tree	0	1	0	0.0000
384	came	0	1	0	0.0000
385	okay	0	0	1	0.2263
386	ssaid	0	1	0	0.0000
387	smacked	0	1	0	0.0000

	Unnamed: 0	neg	neu	pos	compound
388	head	0	1	0	0.0000
389	stop	1	0	0	-0.2960
390	cry	1	0	0	-0.4767
391	tied	0	1	0	0.0000
392	eagle	0	1	0	0.0000
393	stuck	1	0	0	-0.2500
394	stabbed	1	0	0	-0.4404
395	slice	0	1	0	0.0000
396	nut	0	1	0	0.0000
397	sac	0	1	0	0.0000
398	bit	0	1	0	0.0000
399	sucked	1	0	0	-0.4588
400	poked	0	1	0	0.0000
401	hardly	0	1	0	0.0000
402	broke	1	0	0	-0.4215
403	skin	0	1	0	0.0000
404	beside	0	1	0	0.0000
405	go	0	1	0	0.0000
406	thanks	0	0	1	0.4404
407	found	0	1	0	0.0000
408	stay	0	1	0	0.0000
409	way	0	1	0	0.0000
410	trun	0	1	0	0.0000
411	made	0	1	0	0.0000
412	pretty	0	0	1	0.4939
413	disbursted	0	1	0	0.0000
414	drove	0	1	0	0.0000
415	nite	0	1	0	0.0000
416	distributed	0	1	0	0.0000
417	parts	0	1	0	0.0000
418	every	0	1	0	0.0000
419	turned	0	1	0	0.0000
420	wanting	0	1	0	0.0000
421	wouild	0	1	0	0.0000
422	later	0	1	0	0.0000
423	ready	0	0	1	0.3612
424	roughted	0	1	0	0.0000

FIGURA 23: BAG OF WORDS CON ANALISIS DE SENTIMIENTO PARTE 3. FUENTE: ELABORACIÓN PROPIA .

6.1.3.1 *Pie de Sentimiento*

Se ocuparán los resultados de **Bag Of Words** con Análisis de Sentimiento para crear una gráfica tipo pie, representando que tanto de la conversación se tornó en un contexto negativo, positivo o neutral, ver Figura 24.

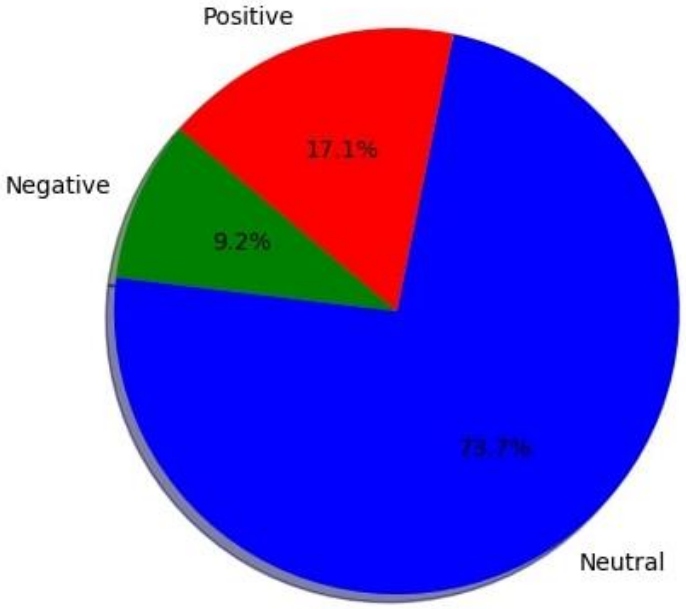

FIGURA 24: PIE DE SENTIMIENTO. FUENTE: ELABORACIÓN PROPIA.

6.1.3.2 *Similarity*

Se realizo una tabla comparando las palabras obtenidas de la conversación con una base de datos con palabras que se usan en un contexto pedófilo, ver Figura 25.

- La fila con la etiqueta "**Unamed:0**" representa las palabras que se están procesando.
- La fila con la etiqueta "**Match**" representa la palabra que se está comparando contra la que se encuentra en "Unamed:0" .
- La fila con la etiqueta "**Similarity**" representa el porcentaje de similitud que tiene la palabra en "Unamed:0" y en "Match".
- La fila con la etiqueta "**Distance**" si las palabras en "Unamed:0" y en "Match" son iguales o no, "0" es no igual y "1" es igual.
- La fila con la etiqueta "**Dis-nomal**" representa "Distance normalized", la cual sirve para saber el valor entre "0" a "1" de igualdad entre "Unamed:0" y "Match", donde entre más alto el valor, más similitud se encuentra.

	Unnamed: 0	Match	Similarity	Distance	Dis-nomal
0	anal	anal	1.000000	0	0.000000
1	ass	ass	1.000000	0	0.000000
2	condoms	condom	0.857143	1	0.142857
3	fuck	fuck	1.000000	0	0.000000
4	love	love	1.000000	0	0.000000
5	making	mating	0.833333	1	0.166667
6	put	put	1.000000	0	0.000000
7	spread	spread	1.000000	0	0.000000

FIGURA 25: SIMILARITY. FUENTE: ELABORACIÓN PROPIA.

6.1.3.3 Possible Similarity

Se realizo una tabla comparando las palabras obtenidas de la conversación con una base de datos con palabras que se usan potencialmente en un contexto pedófilo, ver Figura 26.

- La fila con la etiqueta "**Unamed:0**" representa las palabras que se están procesando.
- La fila con la etiqueta "**Match**" representa la palabra que se está comparando contra la que se encuentra en "**Unamed:0**".
- La fila con la etiqueta "**Similarity**" representa el porcentaje de similitud que tiene la palabra en "**Unamed:0**" y en "**Match**" .
- La fila con la etiqueta "**Distance**" si las palabras en "**Unamed:0**" y en "**Match**" son iguales o no, "**0**" es no igual y "**1**" es igual.
- La fila con la etiqueta "**Dis-nomal**" representa "**Distance normalized**", la cual sirve para saber el valor entre "**0**" a "**1**" de igualdad entre "**Unamed:0**" y "**Match**", donde entre más alto el valor, más similitud se encuentra.

	Unnamed: 0	Match	Similarity	Distance	Dis-nomal
0	animals	animal	0.857143	1	0.142857
1	bit	bit	1.000000	0	0.000000
2	bite	bite	1.000000	0	0.000000
3	girl	girl	1.000000	0	0.000000
4	go	go	1.000000	0	0.000000
5	head	head	1.000000	0	0.000000
6	kid	kid	1.000000	0	0.000000
7	line	line	1.000000	0	0.000000
8	motel	motel	1.000000	0	0.000000
9	queens	queen	0.833333	1	0.166667
10	take	take	1.000000	0	0.000000
11	want	want	1.000000	0	0.000000

FIGURA 26: POSSIBLE SIMILARITY. FUENTE: ELABORACIÓN PROPIA.

6.2. Directorio imagesSimilarity

Directorio donde se guardan solo las imágenes obtenidas por los archivos *xlsx* de "similarity" y "possibleSimilarity "de cada agresor, dentro de este directorio se encuentran otras dos carpetas, "pederastWords" y "possiblePederastWords".

6.2.1 PederastWords

Se crea un archivo PDF por cada agresor, este archivo contiene una tabla donde se muestras la tabla de "similarity" del agresor en particular.

6.2.2 possiblePederastWords

Se crea un archivo PDF por cada agresor, este archivo contiene una tabla donde se muestras la tabla de "possibleSimilarity" del agresor en particular, ver Figura 27.

	Unnamed: 0	Match	Similarity	Distance	Dis-nomal
0	animals	animal	0.857143	1	0.142857
1	bit	bit	1.000000	0	0.000000
2	bite	bite	1.000000	0	0.000000
3	girl	girl	1.000000	0	0.000000
4	go	go	1.000000	0	0.000000
5	head	head	1.000000	0	0.000000
6	kid	kid	1.000000	0	0.000000
7	line	line	1.000000	0	0.000000
8	motel	motel	1.000000	0	0.000000
9	queens	queen	0.833333	1	0.166667
10	take	take	1.000000	0	0.000000
11	want	want	1.000000	0	0.000000

FIGURA 27: POSSIBLE SIMILARITY. FUENTE: ELABORACIÓN PROPIA.

6.3. Directorio ARFF

En este directorio se crean los archivos que utilizará *WEKA* para su análisis en la herramienta, estos archivos se componen de los datos obtenidos en los archivos tipo *xlsx* llamados "PossibleSimilarity" y "Similarity", estos datos son procesados de tal manera para crear los archivos de tipo ".arff"

6.3.1 similarityARFF.arff y similarityPossibleARFF.arff

Documentos con extensión *ARFF* para su análisis en la herramienta *WEKA* tienen el siguiente formato:

- Relation: Nombre del archivo que se realizará.

- Atributo 1: Similarity
  Tipo de Atributo: real
  Descripción: Valor el cual refleja que tan similar es la palabra al momento del procesamiento

- Atributo 2: Normalized_Distance
  Tipo de Atributo: real
  Descripción: Distancia entre las palabras que se estaban comparando utilizando valores más exactos como distancia máxima, distancia mínima y otros datos.

- **Atributo 3**: Apparances
  Tipo de atributo: numérico
  Descripción: Numero de apariciones de la palabra procesada.

- **Atributo 4**: Distance
  Tipo de atributo: 0 o 1
  Descripción: Valor del costo total de transformar la palabra base a la palabra que se está comparando.

  El archivo presenta los campos ya mencionados anteriormente, ver Figura 28.

```
1   @relation Similarity
2   @attribute Similarity real
3   @attribute Nomalized_Distance real
4   @attribute Appearances numeric
5   @attribute Distance {0,1}
6   @data
7   1,0,1,0
8   1,0,4,0
9   0.8571428571428572,0.1428571428571428,1,1
```

FIGURA 28: ARCHIVO ARFF. FUENTE: ELABORACIÓN PROPIA.

6.4. Archivos ARFF con WEKA

Los archivos ARFF se realizaron con el fin de mostrar la eficiencia de la propuesta. Para obtener estos datos se utilizará la plataforma WEKA con el algoritmo J48, los datos que se desean obtener son una matriz de confusión, la *presicion*, el *recall* y el *f-mesure.*
Los archivos ARFF se crearon con 45 conversaciones de la página de Perverted Justice.

6.4.1 WEKA

La plataforma de software de aprendizaje automático y minería de datos (minería de datos) *WEKA* (Waikato Environment for Knowledge Analysis), escrita en *Java* y desarrollada en la Universidad de Waikato la cual se encuentra en Nueva Zelanda.

Inicialmente, la versión original de *WEKA* se desarrolló con una interfaz *TCL/TK* para la modelización de algoritmos implementados en otros lenguajes de programación. También se desarrollaron varias utilidades de preprocesamiento de datos en el lenguaje conocido como *C* para realizar experimentos de aprendizaje automático.

La plataforma *WEKA* se caracteriza por los siguientes parámetros:

- **Disponibilidad**: Esta plataforma de software es gratuita gracias a la Licencia Pública General *GNU*.
- **Personalizable**: Implementado en el lenguaje *Java*, haciéndolo compatible con casi todas las plataformas.
- **Características**: Consta de un amplio repertorio de técnicas de modelado y preprocesamiento de datos.
- **Sencillo**: Muy fácil de operar gracias a la interfaz gráfica de usuario [38].

6.4.2 Algoritmo J48

El algoritmo *C4.5* es un algoritmo de clasificación el cual produce arboles de sedición basado en la teoría de la información. Es una extensión de la versión temprana del algoritmo *ID3* de Ross Quinlan también conocido en *WEKA* como *J48*, la letra J se refiere a *Java*. La implementación *J48* del algoritmo *C4.5* tiene muchas implementaciones adicionales incluyendo el número de apariciones de valores perdidos, arboles de decisión comprimidos, valores de rangos de atributos continuos, derivación de las reglas, entre otros. En la herramienta de minería de datos de *WEKA*, la implementación *open-source* en *Java* de *J48* permite la clasificación ya sea por los árboles de decisión o las reglas generadas por estos mismos [40].

6.4.3 Matriz de confusión

Una matriz de confusión es una matriz o tabla que proporciona información sobre la precisión de un algoritmo de clasificación al clasificar un conjunto de datos, ver Figura 29.
Por tanto, las matrices de confusión son un método para evaluar el rendimiento de los algoritmos de clasificación.

Para problemas de clasificación binaria (me gusta/no me gusta, verdadero/falso, 1/0), la matriz de confusión proporciona cuatro valores de cuadrícula.

Usemos un ejemplo de un sistema que predice a personas que les guste la pizza.

- Los **verdaderos positivos** (TP) representan personas a las que les gusta la pizza y el modelo las clasificó correctamente.
- Los **verdaderos negativos** (TN) representan personas a las que no les gusta la pizza y el modelo las clasificó correctamente.
- Un **falso positivo** (FP) representa a alguien a quien no le gusta la pizza (negativo), pero el clasificador predijo que le gusta la pizza (falso positivo). El FP también se conoce como error de tipo I.
- Un **falso negativo** (FN) representa a alguien a quien le gusta la pizza (un positivo), pero el clasificador predice que no le gusta la pizza (un falso negativo). FN también se conoce como error de tipo II [39].

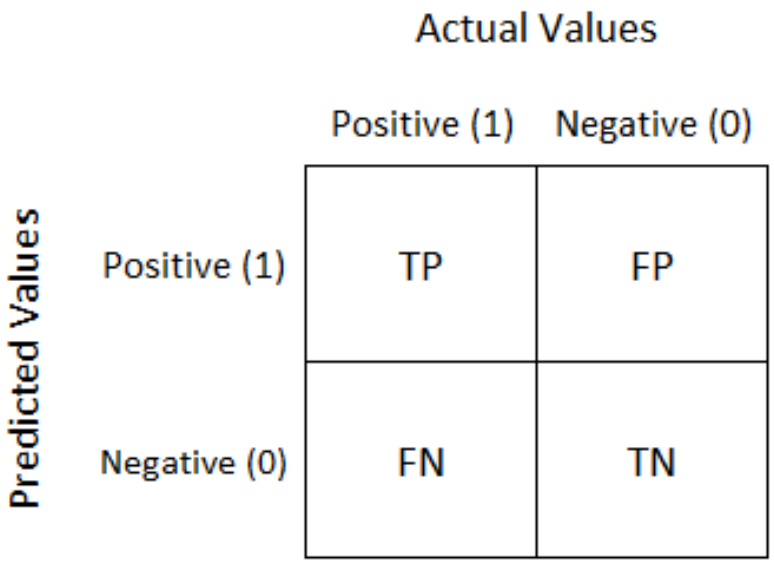

FIGURA 29: MATRIZ DE CONFUSIÓN [41].

6.4.4 Presicion, recall y f-mesure

- La presicion es la probabilidad de que los elementos recuperados seleccionados al azar se consideren relevantes.

- El *recall* es la probabilidad de encontrar un documento relevante seleccionado al azar durante una búsqueda.

- F-mesure se refiere es una medida de la precisión de la prueba. A partir de esto se determina un valor de ponderación único para la precisión y la integridad.

6.4.5 Clasificación

Al clasificar los archivos "similarityARFF.arff" y "similarityPossibleARFF.arff" en *WEKA* con el algoritmo *J48*, los resultados fueron los siguientes:

6.4.5.1 similarityARFF.arff

- Matriz de confusión

La información que nos muestra la matriz de confusión no encontró ningún caso de Falso Positivo y Falso Negativo. Esto se verá reflejado en los datos posteriores, ver Figura 30.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 219   0 |   a = 0
   0  97 |   b = 1
```

FIGURA 30: MATRIZ DE CONFUSIÓN. FUENTE: ELABORACIÓN PROPIA.

- Presicion, recall y f-mesure

Todos los datos nos comparten un resultado, el modelo en caso de las palabras que se utilizan en contexto pedófilo se considera perfecto, sin ningún falso, esto ya sea en la *presicion*, *recall* o *f-mesure*, ver Figura 31.

	Precision	Recall	F-Measure	Class
	1.000	1.000	1.000	0
	1.000	1.000	1.000	1
Weighted Avg.	1.000	1.000	1.000	

FIGURA 31: PRESICION, RECALL Y F-MESURE. FUENTE: ELABORACIÓN PROPIA.

6.4.5.2 similarityPossibleARFF.arff

Matriz de confusión

La información que nos muestra la matriz de confusión nos dice que se encontraron 313 Verdaderos positivos, 0 Negativos Positivos, 1 Falso Positivo y 0 Falsos Negativos, ver Figura 32.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 313   0 |   a = 0
   1  63 |   b = 1
```

FIGURA 32: MATRIZ DE CONFUSIÓN 2. FUENTE: ELABORACIÓN PROPIA.

Presicion, recall y f-mesure

En la clase 0 se visualizan los valores de 0.997, 1.000 y 0.998 respectivamente, el modelo nos dice que tanto la *presicion* como la medida *f-mesure* no son perfectas, sin embargo, tienen valores muy positivos, mientras que el valor de *recall* es perfecto con un valor de 1.000.

En la clase 1 los valores son 0.997 para todas las medidas, lo cual refleja y resalta los valores que se obtuvieron en la matriz de confusión, ver Figura 33.

	Precision	Recall	F-Measure	Class
	0.997	1.000	0.998	0
	1.000	0.984	0.992	1
Weighted Avg.	0.997	0.997	0.997	

FIGURA 33: PRESICION, RECALL Y F-MESURE 2. FUENTE: ELABORACIÓN PROPIA.

Capitulo 7. CONCLUSIONES Y TRABAJO A FUTURO

Las conclusiones de este trabajo de tesis son las siguientes:

Se desarrollo una nueva metodología para la detección de pederastas en un chat, utilizando diferentes líneas de estudio (Psicología, Lingüística, Computación, Criminología) y tecnologías y algoritmos dentro del Procesamiento de Lenguaje Natural.

La metodología que se desarrolló en esta tesis genero resultados de las 45 conversaciones identificadas y analizadas de la página *Perverted Justice*, se recolectaron 14307 palabras en total.

De las 45 conversaciones solo 22 conversaciones fueron capaces de aplicar *web scrapping* debido a formatos diferentes de las etiquetas de *HTML* en cada conversación.

De las palabras recolectadas que se procesaron dieron como resultado que 313 palabras se consideraron de carácter pederasta se creó un *corpus* lingüístico, en donde el mínimo de palabras fue de 1 palabra y el máximo de 38 palabras, con una media de 14.2 palabras en cada conversación.

376 palabras se consideraron de posible carácter pederasta, un mínimo de 3 palabras y un máximo de 49 palabras, con una media de 17.09 palabras en cada conversación.

Las matrices de confusión mostraron resultados muy positivos con solo 1 falso negativo, y esto se refleja con los datos de *presicion, recall* o *f-mesure* con un mínimo de 99.7% y un máximo de 100% de asertividad utilizando el algoritmo *J48* en *WEKA*.
Las implicaciones del futuro sobre el programa son variadas, en primer lugar, la portabilidad del programa es algo que se consideró desde el principio, desde ese punto la idea de realizar un archivo de *Docker* se volvió muy factible, además de cómoda, ya que no se necesitaría instalar demasiadas librerías y sería mucho más sencillo la implementación de este programa.

Los datos que se utilizan para su análisis son conversaciones de la página *Perverted Justice*, el siguiente paso sería que las conversaciones de aplicaciones o servicios de mensajería instantánea como *Messenger*, *WhatsApp*, *Discord* fueran utilizadas para el análisis.
Se puede considerar otra forma de implementación la creación de una extensión para navegadores web como *Chrome*.

La base de datos con palabras que se utilizan en un contexto pederasta por el momento es estática, esto se refiere a que nunca se añaden más palabras y esto nos lleva a la necesidad de automatizar el crecimiento de la base de datos añadiendo más palabras, lo cual es posible con la implementación de aprendizaje supervisado y no supervisado, esto nos ayudará a la mejora del algoritmo además de añadir nuevas formas de análisis de conversaciones.

Capitulo 8. COLABORADORES

A continuación, se agradece a las siguientes personas por el aporte y colaboración en la aplicación, así como en la teoría que se aplicó.

Ariadna Maricel Covarrubias López por su colaboración en el área de lingüística y letras, realizar el análisis de resultados de herramientas léxicas con ejemplos reales y del proceso de análisis de sentimientos, además de la creación de la base de datos con contexto de pedófilo.

Carlos López Santa María por la investigación e implementación del módulo de *Web Scrapping* para la página *Perverted Justice*.

Capitulo 9. BIBLIOGRAFÍA

1. Llonch, E. (2021, 25 mayo). ¿Qué son las redes sociales y cuáles son las más importantes? Recuperado 13 de agosto de 2023, de https://www.cyberclick.es/numerical-blog/que-son-las-redes-sociales-y-cuales-son-las-mas-importantes

2. Florentín, B., & Florentín, B. (2022). Mensajería instantánea. MundoCuentas. https://www.mundocuentas.com/mensajeria-instantanea/

3. Mantener seguros a niñas, niños y adolescentes en internet. (s.f.). UNICEF. https://www.unicef.org/mexico/mantener-seguros-ni%C3%B1as-ni%C3%B1os-y-adolescentes-en-internet

4. Assistant Secretary for Public Affairs (ASPA). (2022, June 30). ¿Qué es el acoso? StopBullying.gov. https://espanol.stopbullying.gov/acoso-escolar-mkb6/qu%C3%A9-es-el-acoso

5. Assistant Secretary for Public Affairs (ASPA). (2022, June 30). ¿Qué es el acoso? StopBullying.gov. https://espanol.stopbullying.gov/acoso-escolar-mkb6/qu%C3%A9-es-el-acoso

6. De Deusto, U. (s.f.). *Negobot*. Flickr. https://www.flickr.com/photos/deusto/10801662136

7. Plaza, V. (2016, 8 diciembre). Un software de inteligencia artificial identifica redes de pederastia en internet. Valencia Plaza. Recuperado 2 de diciembre de 2023, de https://valenciaplaza.com/un-software-de-inteligencia-artificial-identifica-redes-de-pederastia-en-internet

8. Bagnato, J. B. (2022). Procesamiento del Lenguaje Natural (NLP) | Aprende Machine Learning. Aprende Machine Learning. https://www.aprendemachinelearning.com/procesamiento-del-lenguaje-natural-nlp/

9. Simple Bolsa de palabras, (o Bag of Words). (s.f.). http://rstudio-pubs-static.s3.amazonaws.com/268824_161580c9cae441cf85adb95122ee659e.html

10. Brownlee, J. (2019). A gentle introduction to the Bag-of-Words model. MachineLearningMastery.com. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

11. ¿Qué es el análisis de sentimiento? | Una guía completa del análisis de sentimiento. (s.f.). Elastic. https://www.elastic.co/es/what-is/sentiment-analysis

12. De La Iglesia, E. D. (s.f.). Análisis de sentimiento. https://www.campusbigdata.com/big-data-blog/item/158-analisis-de-sentimiento

13. Rédac, T. (2022, August 2). ¿Cómo generar un Wordcloud con Python? Formation Data Science | DataScientest.com. https://datascientest.com/es/como-generar-un-wordcloud-con-python

14. ¿Qué es Python? - Explicación del lenguaje Python - AWS. (s.f.). Amazon Web Services, Inc. Recuperado 2 de septiembre de 2023, de https://aws.amazon.com/es/what-is/python/

15.León, E. (2020, 16 diciembre). Procesamiento del lenguaje natural (PLN) con Python - BAOSS. BAOSS. https://www.baoss.es/procesamiento-del-lenguaje-natural-pln-con-python/

16. Castellanos, E. (2021). Git vs GitHub – ¿Qué es el Control de Versiones y Cómo Funciona? freeCodeCamp.org. https://www.freecodecamp.org/espanol/news/git-vs-github-what-is-version-control-and-how-does-it-work/

17. ¿Qué es el web scraping? (2020, October 9). IONOS Digital Guide. https://www.ionos.mx/digitalguide/paginas-web/desarrollo-web/que-es-el-web-scraping/

18. Datademia. (2021). ¿Qué es web scraping? Datademia. https://datademia.es/blog/que-es-web-scraping

19. Tena, M. (2023, January 19). ¿Qué es la metodología "agile"? BBVA NOTICIAS. https://www.bbva.com/es/innovacion/metodologia-agile-la-revolucion-las-formas-trabajo/

20. Apd, R. (2022). Cómo aplicar la metodología Scrum y qué es el método Scrum. APD España. https://www.apd.es/metodologia-scrum-que-es/

21. Calvo, D. (2019). Metodología SCRUM (Metodología ágil). Diego Calvo. https://www.diegocalvo.es/metodologia-scrum-metodologia-agil/

22. Barbadillo, D. (2021). Stop words. Idital. https://idital.com/diccionario-seo/stop-words/

23. colaboradores de Wikipedia. (2022). Perverted-Justice.com. Wikipedia, La Enciclopedia Libre. https://es.wikipedia.org/wiki/Perverted-Justice.com

24. J2logo. (2022, January 16). Web scraping con Python. Guía de inicio de Beautiful Soup. J2LOGO. https://j2logo.com/python/web-scraping-con-python-guia-inicio-beautifulsoup/

25. Programación en Python - 3. Natural Language Toolkit. (s.f.). https://sites.google.com/view/programacion-en-python/home/3-natural-language-toolkit

26. Alberca, A. S. (2022, May 12). La librería Numpy | Aprende con Alf. Aprende Con Alf. https://aprendeconalf.es/docencia/python/manual/numpy/

27. Coder, R. (2022). Wordclouds (nubes de palabras) en Python. PYTHON CHARTS | Visualización De Datos Con Python. https://python-charts.com/es/ranking/wordcloud-matplotlib/

28. Haosulab. (s.f.). MPlib/README.md at main · haosulab/MPlib. GitHub. https://github.com/haosulab/MPlib/blob/main/README.md

29. Remolino. (2023). Análisis de sentimientos en textos utilizando Python y la librería VaderSentiment. remolinator.com. https://remolinator.com/analisis-de-sentimientos-en-textos-utilizando-python-y-la-libreria-vadersentiment/

30. KeepCoding, R. (2022, November 30). ¿Qué es Matplotlib y cómo funciona? | KeepCoding Bootcamps. KeepCoding Bootcamps. https://keepcoding.io/blog/que-es-matplotlib-y-como-funciona/

31. SoldAI. (s.f.). GitHub - SoldAI/hermetrics: Python library for distance and similarity metrics. GitHub. https://github.com/soldai/hermetrics

32. Sobrino, D. C. (2021, 13 diciembre). Métricas de similitud para cadenas de texto. Parte IV: Biblioteca Hermetrics para Python. Medium. https://medium.com/@diego.campos.sobrino/m%C3%A9tricas-de-similitud-para-cadenas-de-texto-parte-iv-biblioteca-hermetrics-para-python-33404f0ddb70#:~:text=Hermetrics%20est%C3%A1%20dise%C3%B1ada%20para%20facilitar,en%20esta%20serie%20de%20art%C3%ADculos

33. GeeksforGeeks. (2022). OS Module in Python with Examples. GeeksforGeeks. https://www.geeksforgeeks.org/os-module-python-examples/

34. Shutil — Operaciones de archivos de alto nivel. (s.f.). Python documentation. https://docs.python.org/es/3/library/shutil.html

35. Alberca, A. S. (2022, 14 junio). La librería Pandas | Aprende con Alf. Aprende con Alf. https://aprendeconalf.es/docencia/python/manual/pandas/

36. PDFKit. (s.f.). https://pdfkit.org/

37.KeepCoding, R. (2023, 1 junio). ¿Cómo funciona SPACY de Python? | KeepCoding Bootcamps. KeepCoding Bootcamps. https://keepcoding.io/blog/como-funciona-spacy-de-python/#Que_es_spaCy

38. España, R. (2020, 9 julio). ¿Qué es Weka y qué tiene que ver con Big Data? ¿Qué es Weka y qué tiene que ver con Big Data? Recuperado 8 de octubre de 2023, de https://agenciab12.mx/noticia/que-es-weka-que-tiene-que-ver-big-data

39. Khanna, N. (2022, 5 enero). J48 Classification (C4.5 Algorithm) in a nutshell - Nilima Khanna - medium. Medium. https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e

40. De Mariposa, P. (2023). ¿Qué es una matriz de confusión en el aprendizaje automático? Geekflare. https://geekflare.com/es/confusion-matrix-in-machine-learning/

41. Izco, F. (2018, 27 noviembre). *Base de datos corporativa de personas*. Recuperado 18 de septiembre de 2023, de https://bookdown.org/f_izco/BDC-POC/metricas.html

Capitulo 10. ANEXO

AVAL DE IMPRESIÓN DE TESIS

FECHA: 30-10-23

A QUIEN CORRESPONDA:

Por este medio notifico que en mi calidad de asesor *le informo que se ha* ***APROBADO*** *la conclusión de la redacción, avalando la estructura, contenido y aportaciones del documento; por lo tanto,* ***AUTORIZO*** *que* el alumno(a) José Luis Jimenez Aguilera *con numero de matrícula* 201521457 *de la* Ingeniera en Ciencias de la Computación *de esta facultad, realice la impresión de la Tesis titulada:* Detección de pederastia en conversaciones digitales con procesamiento de lenguaje natural.

(nombre y firma)
Dr. Luis Enrique Colmenares Guillén
Asesor de tesis de la Facultad de Ciencias de la Computación

Nota: El Asesor debe ser profesor de la FCC.