UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Cuajimalpa

División de Ciencias de la Comunicación y Diseño

Licenciatura en Tecnologías y Sistemas de Información

Proyecto Terminal 3:

Sistema web de apoyo para la identificación automática de evidencia textual en casos de pedofilia

Por:

Angeles López Flores

Asesores:

Dr. Esaú Villatoro Tello

Mtra. Gabriela Ramírez de la Rosa

Mexico, D.F.

15 de septiembre 2019

Índice

1. Introducción

Actualmente, el constante crecimiento de las tecnologías ha permitido a diversas personas tener a su alcance gran cantidad de servicios en línea disponibles desde cualquier dispositivo conectado a internet. Estos servicios van desde encontrar grandes flujos de información, realizar trámites en línea, compras y transacciones, hasta entablar relaciones personales con otros usuarios sin la necesidad de conocerse físicamente.

Este último ha sido posible gracias a múltiples servicios de interacción social y mensajería instantánea, que permiten conectar con facilidad a personas desde distintas partes del mundo. Algunos ejemplos de este tipo de servicios son los sistemas de chats, blogs, y las presentes redes sociales (Facebook, Twitter, Instagram, Snapchat), entre otros.

Estadísticas muestran que la red social líder es Facebook con más de 2,300 millones de usuarios. Seguido de WhatsApp con 1,600 millones de usuarios, Instagram con 1,000 millones, Twitter con 330 millones y Snapchat con 287 millones de usuario activos a nivel mundial en lo que va del año 2019 [1]. Mientras tanto en México, unos 86 millones de usuarios se encuentran activos en Facebook, 22 millones en Instagram, 10.9 millones en Snapchat, y unos 7.22 millones en Twitter [2].

No obstante, a pesar de que estos nuevos paradigmas de comunicación aportan grandes beneficios, y son fáciles de usar, también presentan grandes riesgos ya que gracias al anonimato que éstas ofrecen, ciertas personas hacen uso de las redes sociales para cometer actos ilícitos como el ciberacoso o más específicamente el acoso sexual a menores.

El acoso sexual a menores de edad, realizado por medio de Internet, conocido en inglés como *Grooming* o *Child Grooming*, y es cometido regularmente por personas adultas denominadas pedófilos. Estas personas buscan acercarse a sus víctimas ganándose su confianza y con el tiempo entablar una conexión emocional. Al mismo tiempo, recuperan información específica de la víctima, para después chantajearla o convencerla de tener una relación más íntima o de naturaleza sexual.

Según el MOCIBA (Módulo sobre Ciberacoso) 2017 del INEGI [3], en México de los usuarios que utilizan internet mayores a 12 años, el 16.8 %

declaro haber vivido alguna situación de acoso cibernético, resultando ser mayor el acoso para mujeres (17.7 %) que para los hombres (16 %). Entre los estados con mayor prevalencia, de este tipo de acoso, se encuentran Tabasco, con un 22.1 %, seguido de Veracruz, Zacatecas, Guanajuato, Aguascalientes e Hidalgo. Algunas de las situaciones experimentadas con mayor frecuencia, por la población que ha vivido ciberacoso de acuerdo al MOCIBA 2017 [3], fueron recibir mensajes ofensivos con un 40.1 %, ser contactados mediante identidades falsas con un 31.4 % y recibir llamadas ofensivas con un 27.5 %.

Por otra parte, cifras revelan que en México, al menos uno de cada siete menores de edad han recibido solicitudes sexuales en las redes sociales, por parte de adultos que se hacen pasar por "amigos". En donde el 80 % de las personas que utiliza estas plataformas aceptan solicitudes de desconocidos y el 43 % habla con ellos [1].

Otro dato preocupante, de acuerdo con datos de la Organización para la Cooperación y Desarrollo Económicos (OCDE) [4], México es el primer lugar a nivel mundial en materia de abuso sexual, violencia física y homicidio de menores de 14 años. Aunque no se menciona el uso de Internet para cometer estos delitos, hay que tener en cuenta que prácticas como el *Grooming*, pueden ser utilizadas por acosadores para acechar previamente a sus víctimas, hasta lograr completar sus objetivos.

En consecuencia, existen campañas y fundaciones sin fines de lucro que buscan ayudar a víctimas de acoso sexual y prevenir a la población sobre los peligros que acechan en Internet. Algunas de estas organizaciones son Secret Survivors México[2], INESSPA[3] (Instituto De Estudios Sobre Sexualidad Y Pareja), Asexoría[4], SavetheChildren[5] y Perverted-Justice[6] (Estados Unidos). Sin embargo, a pesar de estos esfuerzos, aun y desafortunadamente en México no existe ninguna ley que tipifique el *grooming*, dejando indefensos a miles de niños y niñas víctimas de este delito. Es por esta

[1] http://imparcialoaxaca.mx/nacional/344114/advierten-acoso-sexual-a-menores-de-edad-en-redes/

[2] https://www.secretsurvivorsmexico.org/

[3] http://www.inesspa.com/

[4] http://asexoria.net/

[5] https://www.savethechildren.mx/

[6] http://www.perverted-justice.com/

razón, que este proyecto busca proporcionar una herramienta capaz de identificar mensajes de acoso sexual en conversaciones de chat, las cuales, puedan ser utilizadas como evidencia incriminatoria en casos de pedofilia.

1.1. Justificación

El *grooming* es una práctica que ha crecido en Internet ya que cada vez son más los casos que se presentan por este tipo de acoso. Sin embargo, solo pocos de ellos son denunciados y de éstos la minoría queda resuelta. Según la Comisión Ejecutiva de Atención a Víctimas (CEAV) [5] en promedio, de cada cien casos de agresiones sexuales que se cometen en México, sólo seis son denunciados y de éstas solo un tercio llegan a ser consignadas ante un juez.

Entre las estrategias más utilizadas por agencias gubernamentales para combatir este tipo de delitos, son la utilización de perfiles ficticios en sistemas de chat, en los cuales, agentes de policía toman el papel de señuelo haciéndose pasar por niños o adolescentes. Esto, con el objetivo de engañar y posteriormente arrestar a los usuarios que intentan acosarlos sexualmente.

Una vez arrestado el agresor los agentes de policía necesitan encontrar evidencias que permitan culpar a una persona de haber cometido *grooming*. Dado que las salas de chat son el principal medio donde se desarrolla este delito, el conjunto de evidencias normalmente se compone de aquellas líneas o mensajes de la conversación que delatan a un usuario como acosador. A esto último, se le conoce como identificación del comportamiento pedófilo.

No obstante, estas soluciones manuales a menudo son rebasadas por la gran cantidad de usuarios pedófilos que se encuentran al acecho de menores. Además de la complejidad que representa la tarea de identificar aquellas líneas acusatorias dentro de una conversación. Es debido a esto, que se ha hecho necesario el desarrollo de herramientas, que puedan servir en la detección automática de evidencia incriminatoria presente en una conversación.

En consecuencia, este trabajo se enfoca en generar una herramienta

automática de apoyo para identificar aquella evidencia que permite incriminar a un pedófilo. Para ello, esta propuesta pretende crear una herramienta computacional que sea capaz de detectar líneas incriminatorias de acoso sexual dentro de una conversación. Asimismo, permita a un experto visualizar de manera clara aquellas intervenciones por parte del usuario acosador, que son más distintivas de actividades pedófilas.

1.2. Objetivos

1.2.1. General

Desarrollar una herramienta computacional que facilite a un experto la identificación de evidencia de acoso sexual en conversaciones de chat.

1.2.2. Específicos

- Analizar las conversaciones de acosadores sexuales para extraer los atributos más relevantes de acoso sexual.
- Utilizar los atributos de acoso identificados en un esquema de aprendizaje supervisado para detectar mensajes más representativos de acoso sexual.
- Crear una herramienta de visualización para el apoyo en la identificación de líneas incriminatorias dentro de una conversación.

1.3. Organización del documento

El resto de este documento se encuentra organizado por los siguientes capítulos:

- Marco Teórico: En esta sección, se explican los conceptos y elementos teóricos que son fundamentales para la comprensión del contexto de este proyecto.

- Trabajo Relacionado: En este apartado, se presenta una revisión de los trabajos previos relacionados a la problemática de identificación de depredadores sexuales y comportamiento depredador.
- Método propuesto: En este capítulo, se describe a detalle el método para la detección de líneas detonantes de acoso sexual, propuesto en este proyecto, así como los experimentos realizados a lo largo de la elaboración del mismo.
- Desarrollo del Sistema Web: Esta sección, consta del esquema general del sistema y los componentes que lo conforman. Además de explicar cada una de las etapas desarrolladas en la construcción del sistema.
- Conclusiones: Por último, se habla de los resultados obtenidos, objetivos logrados y trabajo a futuro de este proyecto.

2. Marco Teórico

El objetivo de esta sección es introducir brevemente al lector algunos conceptos y elementos teóricos que son fundamentales para entender este proyecto terminal. En primer lugar, se introduce al lector el tema de aprendizaje automático, así como también se explican los algoritmos de aprendizaje y las métricas de evaluación. Posteriormente, se pretende introducir al lector a la tarea de clasificación automática de textos, representación de textos y finalmente perfilado de autor.

2.1. Aprendizaje Automático

El Aprendizaje Automático o en inglés *Machine Learning* es una rama de la Inteligencia Artificial, utilizada para darle a las computadoras la capacidad de aprender. Una de las definiciones más usadas de aprendizaje automático, es la siguiente:

Se dice que una computadora es capaz de aprender de una experiencia $\boldsymbol{E}$ con respecto a una tarea $\boldsymbol{T}$ y una medida de desempeño $\boldsymbol{P}$, si su desempeño en la tarea $\boldsymbol{T}$, medida por medio de $\boldsymbol{P}$, mejora con la experiencia $\boldsymbol{E}$ [6].

Con experiencia, se refiere al conjunto de instancias que sirven como ejemplos de entrenamiento. En donde estas instancias pasan por un proceso de selección de atributos y una etapa de representación de documentos.

La selección de atributos tiene como objetivo seleccionar aquellas características dentro de un texto que mejoran el desempeño predictivo de los modelos, haciéndolos más rápidos y menos costosos [7]. Mientras que la representación de documentos tiene como finalidad transformar documentos escritos en lenguaje natural en representaciones adecuadas para la extracción de información [8]. Este concepto se describe más a detalle en la subsección 2.2.1.

En cuanto al aprendizaje automático, existen tres tipos diferentes de aprendizaje, los cuales son:

- Aprendizaje Supervisado: Se basa en aprender a predecir los valores

objetivos a partir de datos etiquetados, es decir, por medio de ejemplos, predecir una clase de un conjunto de clases predefinidas.

- Aprendizaje No Supervisado: Se basa en encontrar una estructura en datos no etiquetados, es decir, encontrar conocimiento o estructuras útiles para agrupar los documentos en grupos similares.
- Aprendizaje Semi-supervisado: Se basa en aprender a partir de un conjunto de datos, de los cuales solo una pequeña parte de los datos están etiquetados, mientras que la otra parte no lo está.

2.1.1. Algoritmos de Aprendizaje

Los algoritmos de aprendizaje se agrupan de acuerdo con el tipo de aprendizaje que se requiere (supervisado o no supervisado). Sin embargo, en este proyecto terminal se trabajó únicamente con algoritmos de aprendizaje supervisado, algunos de estos algoritmos se describen a continuación:

- Bayes Ingenuo (Naïve Bayes): Es un clasificador probabilístico fundamentado en el teorema de Bayes, el cual, ayuda a inferir la probabilidad de pertenencia a una clase. Es decir, la probabilidad de que una muestra dada pertenezca a una clase en particular [9]. Este clasificador asume la independencia de los atributos entre las diferentes clases del conjunto de entrenamiento y es altamente eficiente para aprender y predecir [10].
- Máquina de Vectores de Soporte (SVM): Es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria. Este método gira en torno a la noción de un "margen", en donde cada lado del hiperplano separa los elementos en dos clases diferentes [11].
- Clasificador K-vecinos (k-Nearest Neighbors): La idea básica de este clasificador es que una nueva muestra sin etiqueta se va a clasificar en la clase más frecuente a la que pertenecen sus K-vecinos más cercanos [9].
- Árbol de decisión: Este clasificador se basa en la construcción de un árbol en donde cada nodo interno es un atributo de la instancia, y

cada rama representa un posible valor de ese atributo, finalmente cada hoja específica el valor de la categoría (decisión) [6].

- Bosque aleatorio (Random Forest): Es una técnica que se basa en un conjunto de árboles de decisión como predictores construidos de forma individual, de los cuales se toma una decisión final mediante un voto plural entre los mismos predictores [12].

2.1.2. Métricas de Evaluación

Las métricas de evaluación son utilizadas para medir el grado de confiabilidad que tiene un modelo de aprendizaje automático.

Una herramienta que permite visualizar el desempeño de predicción de un algoritmo empleado es la matriz de confusión. Las columnas de la matriz representan el número de predicciones de cada clase, mientras que las filas representan el número de predicciones correctas para cada clase. La figura 1 muestra un ejemplo de este concepto.

	Predicciones de clase positiva	**Predicciones de clase negativa**
Instancias reales positivas	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Instancias reales negativas	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Figura 1: Ejemplo de matriz de confusión.

Las métricas de evaluación más utilizadas para medir el desempeño de algoritmos de aprendizaje supervisado son:

- Precisión: Fracción de las predicciones positivas que son correctas.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

En donde:
TN = Cantidad de predicciones negativas que realmente son negativas.
TP = Cantidad de predicciones positivas que realmente son positivas.
FN = Cantidad de predicciones negativas que en realidad no son negativas.
FP = Cantidad de predicciones positivas que en realidad no son positivas.

- Exhaustividad (Recall): Fracción de las instancias positivas que el clasificador identifica correctamente como positiva.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

- Medida F-Score: Es una combinación de las medidas de precisión y recall en un solo valor, el cual permite tener un equilibrio entre los resultados de cada medida individual.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{3}$$

Una técnica muy utilizada para la evaluar los resultados en un modelo es la validación cruzada. Esta técnica consiste en dividir los datos en varios segmentos principales. Un conjunto de estos segmentos es utilizado para entrenar un modelo, mientras que el otro es utilizado para evaluar el mismo modelo. Estos conjuntos de entrenamiento y evaluación deben cruzarse en rondas sucesivas de modo que cada conjunto de datos tenga la misma posibilidad de ser validado [13].

La forma básica de validación cruzada es la validación cruzada *k-fold*. En la cual, los datos se dividen primero en k segmentos o pliegues de igual tamaño (o casi igual). Posteriormente, se realizan k iteraciones de entrenamiento y validación de tal manera que dentro de cada iteración se retiene un pliegue diferente de los datos para la validación mientras que los k pliegues restantes se usan para el aprendizaje [13].

La figura 2 muestra un ejemplo de validación cruzada con k = 3. En donde las secciones más oscuras son para el entrenamiento, mientras que las más claras son para la validación.

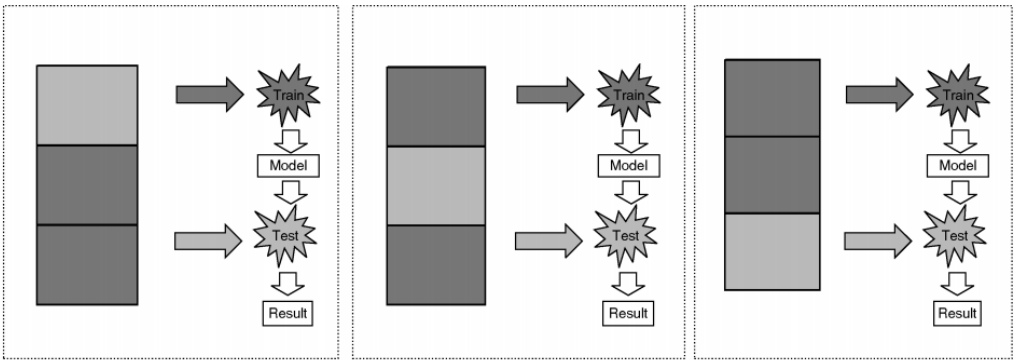

Figura 2: Ejemplo de validación cruzada con tres pliegues. Figura tomada de [13].

2.2. Clasificación Automática de Textos

La clasificación automática de textos es una tarea general de aprendizaje automático que consiste en asignar a un documento una etiqueta de una clase predefinida [8]. Esto significa, que requiere de un conjunto de documentos previamente clasificados que sirvan como ejemplos, de tal manera que el clasificador automático genere una clasificación propia frente a un documento desconocido [14].

Una tarea que se trabaja en la clasificación de textos es el perfilado de autor. El cual, tiene como objetivo predecir si un texto pertenece o no a un grupo de autores que comparten ciertas características; como el género, la edad, el nivel educativo, la región geográfica, entre otros [15, 16].

Primeramente, para realizar una tarea de clasificación automática de textos, se deben obtener los atributos que describan el texto a clasificar. Así como también transformarlos a una representación adecuada para ser utilizados por algoritmos de aprendizaje automático. En la siguiente sección, se explica con mayor detalle algunos de los métodos más utilizados

para representar textos.

2.2.1. Representación de Textos

Algunos de los métodos más utilizados para representar textos son los siguientes:

- Bolsa de Palabras (BoW): Denominado BoW por sus siglas en inglés *Bag of Words*, es la representación más común en los sistemas de clasificación de textos. En este modelo un documento es representado por medio de un vector numérico que indica el peso de la ocurrencia de las palabras del vocabulario de la colección de dicho documento. Sin embargo, con este método los documentos pierden contexto ya que en esta representación no se sigue el mismo orden de aparición de los términos. A continuación, en la figura 3 se ilustra el modelo antes descrito, con el texto: *Compré pocas copas*, *pocas copas compré*, *como compré pocas copas*, *pocas copas pagaré* [8].

$\mathcal{V} =$	{compré,	pocas,	copas,	como,	pagaré}
$\vec{d_i} =$	(3,	4,	4,	1,	1)

Figura 3: Representación de documentos utilizando BOW [8].

En el ejemplo anterior, el tamaño del vector está determinado por el tamaño del vocabulario del texto en cuestión. Mientras que los números de cada término en el vector representan la frecuencia de esa palabra en el texto.

- N-gramas: Este enfoque consiste en utilizar secuencias de elementos (tokens) que se solapan, los elementos pueden ser palabras o caracteres. Los n-gramas resultan ser muy útiles en tareas de clasificación ya que pueden capturar información estilográfica. En el caso de n-gramas de caracteres e información sintáctica en el caso de n-gramas de palabras. Además de que son capaces de añadir contexto a la representación. A modo ilustrativo la figura 4 muestra los bigramas (N-grama con N=2) de palabras derivados del siguiente texto: *Compré pocas copas* [8].

$V =$	{compré_pocas,	pocas_copas}
$d_i =$	(1,	1)

Figura 4: Representación de documentos utilizando bigramas de palabras [8].

Para cada uno de los modelos de representación de documentos antes descritos, se utiliza un *"esquema de pesado"* el cual ayuda a calcular la relevancia de cada término dentro de un documento. Algunos esquemas más utilizados son los siguientes:

- Booleano: Consiste en asignar el peso con valor de 1 si la palabra aparece en el documento y 0 en caso contrario.
- Frecuencia del término (tf): Es un esquema que asigna a cada término un valor igual a la cantidad de veces que aparece en el documento.
- Frecuencia Relativa (tf-idf): Este esquema corresponde al producto de dos factores tf e idf, este esquema mide cuánta información provee el término; es decir, la frecuencia de ocurrencia del término en comparación con otros documentos de la colección. Formalmente el valor tf-idf para un término t en el documento d se calcula con la ecuación 4.

$$tfidf_{t,d} = tf_{t,d} \times \log_2\left(\frac{N}{tf_t}\right) \tag{4}$$

En donde $tf_{t,d}$ es el número de ocurrencias de t en d, N es la cantidad total de documentos y *dfi* es el número de documentos que contienen ese término t. En base a la combinación de estos dos factores, el esquema *tf-idf* indica la relevancia de un término de acuerdo con la frecuencia de aparición dentro de la colección [8].

3. Trabajo Relacionado

En esta sección se resumen algunos trabajos relacionados a la problemática de identificar aquellas líneas dentro de una conversación que son más representativas de acoso sexual. Sin embargo, dado que estos trabajos fueron presentados como resultado de la competencia CLEF PAN (Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse) en 2012 [17], se iniciará por una breve descripción de esta competencia, en la subsección de antecedentes.

3.1. Antecedentes

CLEF PAN [7], es un congreso donde se abordan temas relacionados a la atribución de autoría, plagio de textos y vandalismo informático. En el año 2012 se incluyó dentro de atribución de autoría la subtarea "Sexual Predator Identification" en la cual, se presentan dos tareas a resolver: 1) Identificación de usuarios pedófilos y 2) Identificación de líneas pedófilas. La primera, está enfocada en la identificación de usuarios pedófilos que se encuentran presentes dentro de una conversación. Mientras que, para la segunda tarea, se pretende identificar aquellas líneas de conversación que son más representativas de acoso sexual y que puedan ser utilizadas como evidencia incriminatoria contra un usuario pedófilo.

Para la solución de ambas tareas, se presentaron diversas propuestas de solución por parte de los participantes, de las cuales, para la primera tarea la mejor solución alcanzó un puntaje $F_{0,5}$ de 0.9346 la cual, fue desarrollada por Escalante et al [18]. Por otra parte, para la segunda tarea, la mejor solución fue propuesta por Grozea y Popescu [19], con un puntaje $F_{\beta=3}$ de 0.4762.

Dado que, la propuesta de este proyecto terminal se enfoca en apoyar en la identificación de aquella evidencia dentro de una conversación que permita incriminar a un pedófilo, es que en la siguiente sub-sección se resumen algunas soluciones propuestas en la competencia CLEF PAN 2012 para la tarea de "Identificación de líneas pedófilas".

[7] https://pan.webis.de/clef12/pan12-web/

3.2. Identificación de líneas pedófilas

El trabajo propuesto por Grozea y Popescu [19], alcanzó el primer puesto en la tarea de "Identificación de líneas pedófilas" de la competencia CLEF PAN 2012. El enfoque utilizado en esta solución se basó simplemente en devolver como sospechosas todas las líneas escritas por los usuarios que fueron etiquetados como pedófilos en su solución a la tarea de "Identificación de usuarios pedófilos". Los autores no descartaron ninguna línea escrita por estos usuarios, ya que consideran que cada una de estas líneas de chat tienen un propósito.

Por otra parte, el segundo puesto de la competencia lo obtuvo el trabajo realizado por Kontostathis et al [20]. En este trabajo, los autores utilizaron un software llamado ChatCoder 2.0; el cual fue programado por ellos mismos. Este software fue desarrollado con el objetivo de identificar aquellas publicaciones (líneas) más representativas de acoso sexual en una conversación en línea. ChatCoder 2.0 agrupa las líneas de conversación en tres grupos principales: Intercambio de información personal, preparación (Grooming) y acercamiento (Approach). Para distinguir a que grupo pertenece cada línea de conversación, se apoyaron en una serie de atributos diferentes que se recopilan para cada una de las líneas. Algunos de estos atributos son, el número total de palabras, número de pronombres en primera, segunda o tercera persona, número de palabras referentes a información personal, relaciones o actividades, número de sustantivos referentes a la familia, entre otros. Finalmente, se seleccionaron como resultado para la tarea de "Identificación de líneas pedófilas" todas aquellas líneas de conversación que el software ChatCoder 2.0 agrupo en las categorías: preparación (Grooming) y enfoque (Approach).

El tercer puesto de la competencia fue realizado por Peersman et al [21]. Su propuesta se basó en una división de tres etapas o categorías. La primera de estas categorías: "El tema sexual" se refiere a aquellas discusiones dentro de la conversación, en las que se mencionan partes del cuerpo o actos sexuales. La segunda: "Reencuadre", agrupa a todas aquellas líneas donde se hace mención de actividades como jugar, practicar o enseñar actos sexuales. La tercera y última, se refiere a expresiones que hacen referencia a una reunión en persona. En esta propuesta, los autores seleccionaron como resultado de líneas sospechosas todas aquellas que

se encontraron dentro de alguna de las tres categorías descritas anteriormente. Adicional a esto, también se incluyeron aquellas líneas donde se identificaron temas como la solicitud de datos (fotos, videos o el uso de cámaras web), el aislamiento de la supervisión de un adulto y términos de expresiones en diminutivo (ej. barriguita).

Otra propuesta de solución fue propuesta por Morris y Hirst [22], el cual obtuvo el quinto lugar en la tabla de posiciones de la competencia. En este trabajo los autores se apoyaron de una "lista negra" construida por ellos mismos, la cual contenía términos sexualmente explícitos obtenidos directamente del corpus proporcionado por los organizadores de CLEF PAN 2012. Esta lista, también contenía términos referentes a envíos de fotos (de desnudos) y la programación de citas o encuentros en persona. Además de lo anterior, los autores también consideraron algunas características de comportamiento como el número de iniciaciones (número de veces que el acosador inicia una conversación), numero de preguntas hechas por el usuario acosador y el tiempo que tarda en contestar. Para distinguir entre un usuario acosador y uno que no lo es (primera tarea), los autores formaron un modelo SVM lineal utilizando únicamente características léxicas. Posteriormente, utilizaron los pesos resultantes sobre unigramas y bigramas para inducir una ponderación de "acoso" en todos los términos. De esta forma, se seleccionaron como líneas sospechosas todas aquellas en las que la suma de sus pesos de n-gramas rebasa un umbral establecido manualmente por los mismos autores. Asimismo, también se agregaron todas las líneas que contenían cualquier término de la "lista negra".

Por último, un trabajo que presenta otra solución fue el propuesto por Villatoro Tello et al en [23]. En esta propuesta, los autores consideran que un usuario pedófilo sigue tres etapas principales cuando se acerca a un niño. En la primera, el acosador obtiene acceso a la víctima, en la segunda involucra a la víctima en una relación engañosa, y en la tercera inicia y prolonga una relación de abuso sexual. La solución, planteada por estos autores, fue dividir automáticamente todas las conversaciones donde aparezca un usuario pedófilo en tres secciones; sin embargo, no se considera ningún tipo de frontera textual para esta división, es decir, no se identifica dónde comienza y termina la primera etapa. Posteriormente, se genera un modelo de lenguaje solo de la 2da y 3ra etapa del usuario etiquetado

como acosador sexual. Finalmente se calcula la perplejidad[8] frente al modelo generado de cada una de las líneas escritas por el usuario acosador y se presentan como líneas más representativas de acoso sexual aquellas con menor valor de perplejidad.

Como se mencionó, al inicio de esta sección, los trabajos anteriormente descritos son propuestas de solución resultantes de la competencia CLEF PAN 2012. Sin embargo, la principal diferencia con estos trabajos es el corpus obtenido, ya que, en el caso de la competencia a ninguno de los participantes se les proporciono un corpus para la solución de la segunda tarea "Identificación de líneas pedófilas". Es decir, los participantes debían partir de su solución propuesta en la primera tarea para resolver la segunda, por lo que cada participante utilizó las líneas de usuarios que identificaron como acosadores en su solución propuesta a la tarea uno. En diferencia, este proyecto utilizó el mismo corpus utilizado en la competencia CLEF PAN 2012; sin embargo, la diferencia principal fue que se obtuvo acceso al archivo estándar de solución de la primera y segunda tarea.

La obtención de los archivos de solución, facilitaron la construcción de un corpus etiquetado construido a partir del corpus original, esto se describe más a detalle en la subsección 4.1 de este proyecto terminal. El corpus construido se encuentra compuesto de líneas más representativas de acoso sexual escritas solamente por usuarios acosadores, esto permite manejar el problema con una metodología de clasificación supervisada. Por el caso contrario, en los trabajos anteriormente descritos, el no contar con un corpus etiquetado representaba una dificultad para incluir soluciones basadas en *Machine Learning*, por lo que se aplicaron otras metodologías.

Ciertamente existen algunas ventajas frente a los participantes de CLEF PAN 2012. Sin embargo, estos trabajos inspiran a continuar trabajando sobre esta problemática ya que se espera mejorar los resultados obtenidos en esta competencia, dada las principales ventajas y diferencias con las que se encuentra este proyecto terminal.

[8]Calcula qué tan bien un modelo de probabilidad predice una muestra.

4. Método Propuesto

En el presente capítulo, se describe el método propuesto generado para este proyecto terminal. Este método se basa en una metodología de aprendizaje supervisado, el cual se puede observar en la figura 5.

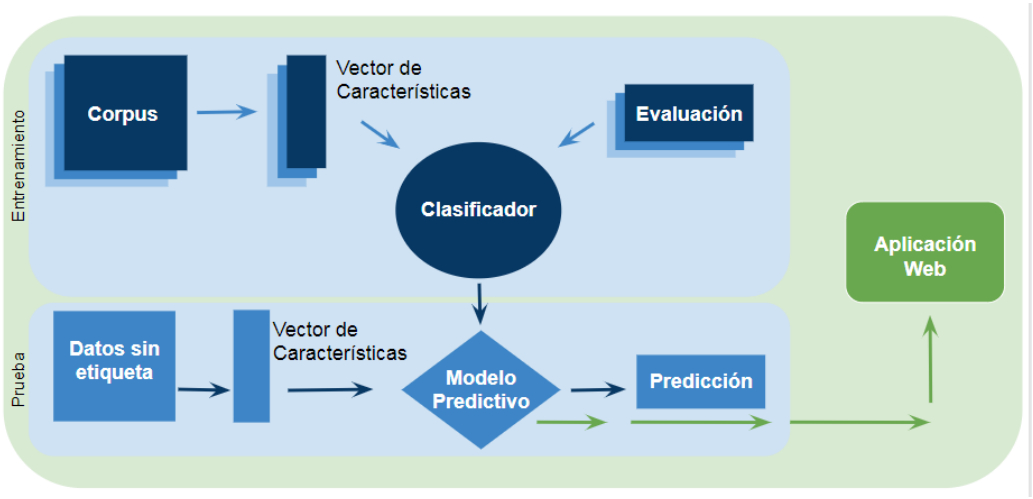

Figura 5: Representación del método propuesto basado en una metodología de aprendizaje supervisado.

En la figura 5 se puede observar que la metodología se divide en dos etapas: entrenamiento y prueba. Para la etapa de entrenamiento, se utiliza un conjunto de textos que sirven como ejemplos para un algoritmo de aprendizaje, el cual es denominado Corpus. Por otra parte, el vector de características se refiere a la representación de textos (concepto definido en la sección 2.2.1) utilizada para representar los atributos identificados en los textos del corpus. El clasificador, se refiere al algoritmo de aprendizaje utilizado para entrenar un modelo de predicción. Mientras que la evaluación, se refiere a la métrica de evaluación utilizada para evaluar el desempeño del modelo entrenado. Los conceptos de algoritmos de aprendizaje y métricas de evaluación, fueron explicados en el Marco Teórico de este proyecto terminal (sección 2).

En la etapa de Prueba, se selecciona el modelo predictivo más óptimo generado, para ser utilizado con datos nuevos denominados en la fi-

gura 5, como datos sin etiquetas. Estos datos, pasan por el mismo proceso de representación de textos utilizada en la etapa de entrenamiento. Posteriormente, pasan por el modelo predictivo seleccionado; para finalmente, obtener la clasificación a la que pertenece el nuevo dato.

Por último, el modelo predictivo seleccionado, es utilizado en la implementación de una herramienta web. Esta última parte se describe en la sección 5 de este proyecto terminal.

A continuación, en las siguientes subsecciones se describe más a detalle el Corpus utilizado, así como la representación de los atributos identificados y por último se explican los experimentos realizados en este proyecto terminal.

4.1. Corpus

El corpus utilizado en este proyecto, fue el utilizado en la competencia CLEF PAN 2012[9], el cual comprende una gran cantidad de conversaciones de chat en inglés que incluyen usuarios pedófilos reales. En la figura 6 podemos ver un ejemplo de la constitución de este corpus.

```
<conversation id="0a1af5368270c9228dcfa59a108cb5ba">
    <message line="1">
      <author>a23878d255460147a5430b03a9c83236</author>
      <time>21:59</time>
      <text>hi</text>
    </message>
    <message line="2">
      <author>220840d2c4fda35d80b9e3855263d7b9</author>
      <time>21:59</time>
      <text>hi</text>
    </message>
    ...
</conversation>
```

Figura 6: Ejemplo del contenido del corpus.

En la figura anterior se puede ver que cada conversación del corpus se encuentra en formato XML y ésta se identifica por el atributo "Id" de la etiqueta "conversation". Asimismo, por cada mensaje de la conversación, se

[9]https://pan.webis.de/clef12/pan12-web/

guarda el autor (identificado por el atributo "Id" de la etiqueta "author"), el horario del mensaje (etiqueta "time") y el texto del mensaje (etiqueta "text").

La totalidad de este corpus está compuesto por dos partes: la primera, denominada "train" es destinada para la etapa de entrenamiento; la segunda parte, denominada "test" es utilizada en la etapa de prueba. De esta forma, se recomienda utilizar la parte "train" para entrenar al modelo que se encuentra desarrollando, mientras que la parte "test" se recomienda utilizar para evaluar el desempeño del modelo ya entrenado. La tabla 1, presenta una comparativa del contenido de la parte "train" y "test" del corpus obtenido.

Contenido	Train	Test
Total de Conversaciones	66,927	155,128
Total de Usuarios	97,689	218,702
Usuarios Pedófilos	142	254
Conversaciones con usuarios pedófilos	1,333	3,737

Tabla 1: Distribución del contenido de la parte "train" y "test" del corpus.

Sin embargo, debido a los objetivos que motivan este proyecto terminal, se utilizó únicamente la parte "test" del corpus descrito anteriormente, ya que esta parte, era la única del corpus que contenía un archivo de referencia a las líneas pedófilas que se encontraban presentes en las conversaciones. Dicho en otras palabras, la parte "test", era la única del corpus que contenía el archivo de solución a la segunda tarea "Identificación de líneas pedófilas" de la competencia CLEF PAN 2012.

Dado lo anterior, se construyó una versión simplificada de la parte "test" del corpus original obtenido de CLEF PAN 2012. Esta versión simplificada, guarda únicamente las conversaciones con usuarios pedófilos, ya que éstas, son las que más interesa analizar. A partir de este punto, se referirá como corpus de conversaciones pedófilas, utilizado en este trabajo, únicamente a la parte de "test" del corpus original (Tabla 1).

A continuación, en la siguiente subsección se presenta una serie de estadísticas realizadas al corpus de conversaciones pedófilas. Esto con el

fin de analizar e identificar aquellas características que pueden ayudar a identificar una línea de chat incriminatoria de acoso sexual.

4.1.1. Estadísticas

Las siguientes estadísticas fueron realizadas con el corpus de conversaciones pedófilas.

1. **Líneas y horarios por conversación:**

   El objetivo de estas estadísticas es analizar la cantidad de líneas que conforman las conversaciones de usuarios pedófilos y en qué rango de horario son más presentes. Para ello se clasificaron los horarios en las siguientes cuatro categorías, de acuerdo con los diferentes momentos que transcurren en el día:

   - Madrugada - 00:00 - 05:59
   - Mañana - 06:00 - 11:59
   - Tarde - 12:00 - 17:59
   - Noche - 18:00 - 23:59

   Posteriormente, por cada conversación se calculó lo siguiente:

   - Número de líneas generadas en la conversación.
   - Categoría de horario, de acuerdo con el rango de horarios presentes en la conversación.

   Las figuras 7 y 8 presentan gráficas de las estadísticas realizadas con estos datos. En la gráfica 7, se puede observar que la mayoría de las conversaciones tienen en promedio 80 líneas de chat, variando en un rango desde 50 hasta 110 líneas. Por otra parte, en la gráfica 8 se puede observar que un gran número de las conversaciones fueron realizadas en la categoría noche (de acuerdo a la clasificación descrita anteriormente). Seguido de la categoría tarde, madrugada y mañana.

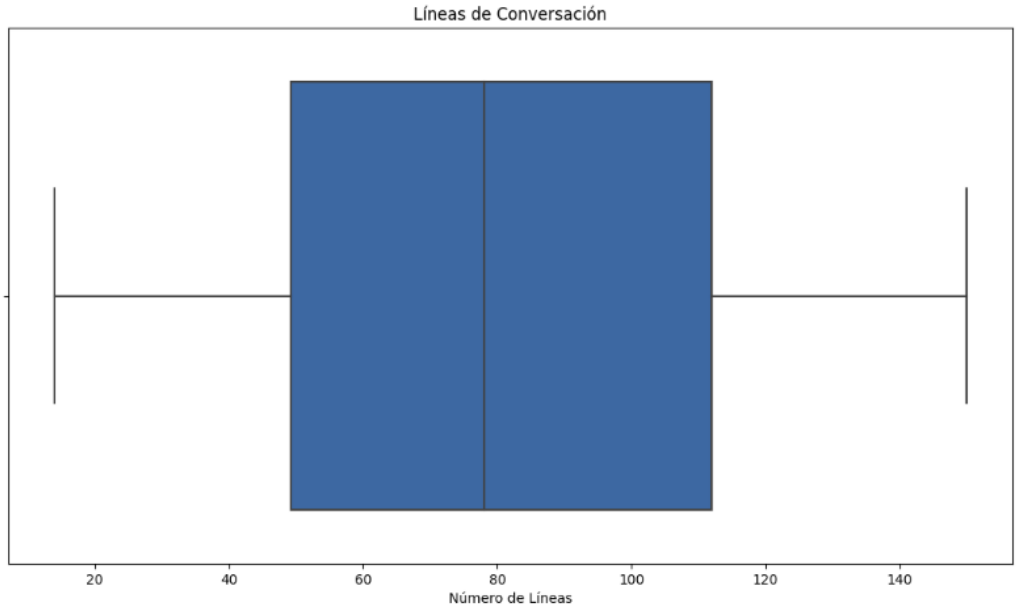

Figura 7: Distribución general del número de líneas de conversación. (Promedio: 80.47 | Desviación Estándar: 37.34)

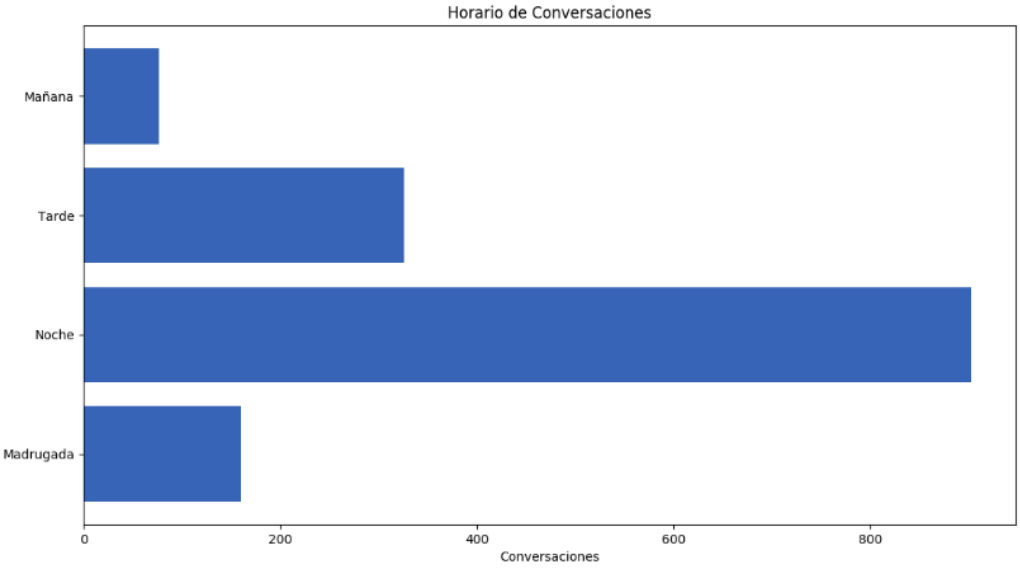

Figura 8: Cantidad de conversaciones por cada categoría de horarios. (Mañana: 77 | Tarde: 326 | Noche: 903 | Madrugada: 160)

2. **Conversaciones por usuario pedófilo y líneas generadas por conversación:**

   Estas estadísticas se realizaron con el objetivo de observar cuánta presencia en conversaciones tienen cada uno de los usuarios pedófilos del corpus. Para ello, se realizó un filtrado de estos usuarios, creando un archivo por usuario pedófilo. Posteriormente, se asignaron cada una de las conversaciones al archivo del acosador correspondiente. De esta forma cada archivo contiene el conjunto de todas las conversaciones en las que participa ese usuario. Una vez obtenido lo anterior, por cada archivo se calculó lo siguiente:

   - Número de conversaciones en las que participa el usuario.
   - Promedio del total de líneas generadas en las conversaciones.
   - Promedio de líneas generadas por el usuario acosador en las conversaciones.
   - Promedio de líneas generadas por la víctima en las conversaciones.

   Las figuras 9, 10, 11 y 12 presentan estadísticas generales de los datos anteriormente descritos. En estas gráficas se puede observar que cada usuario pedófilo del corpus participa en promedio en 6 conversaciones diferentes. De igual forma, el promedio de líneas generadas en estas conversaciones (por usuario pedófilo), es de: 84.79. Adicional a esto, se encuentra que el usuario pedófilo, genera en promedio 43 líneas, mientras que la víctima genera un promedio de 41, ligeramente menor a las del acosador.

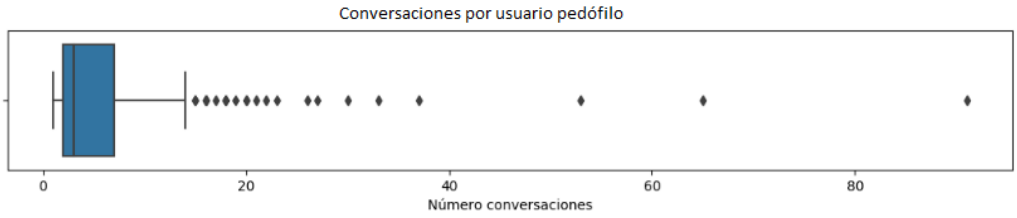

Figura 9: Cantidad de conversaciones en las que participa un usuario pedófilo. (Promedio: 6.48 | Desviación Estándar: 9.71)

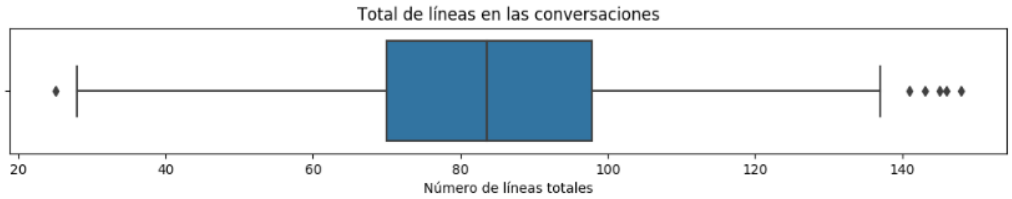

Figura 10: Cantidad de líneas totales generadas en las conversaciones. (Promedio: 84.79 | Desviación Estándar: 24.26)

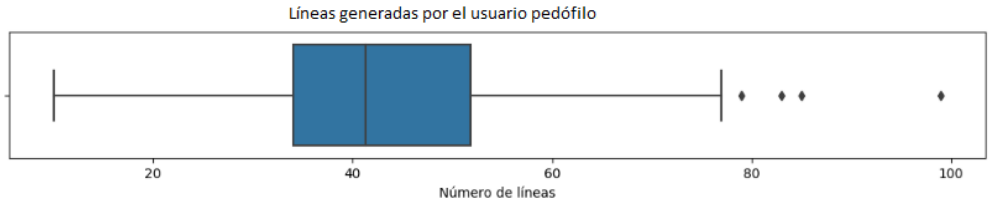

Figura 11: Cantidad de líneas generadas por el usuario acosador en las conversaciones. (Promedio: 43.43 | Desviación Estándar: 14.19)

Figura 12: Cantidad de líneas generadas por la víctima en las conversaciones. (Promedio: 41.36 | Desviación Estándar: 12.90)

3. **Toma de turnos entre usuarios por conversaciones:**

   El objetivo de estas estadísticas es observar cómo se distribuyen los turnos de los usuarios en la conversación (víctima y acosador). Para ello, se tomaron en cuenta dos cuestiones:

   - Interacciones en la conversación: Se refiere al conjunto de líneas de un usuario hasta que el otro toma el turno.
   - Líneas en la conversación: Se refiere a cada línea escrita por un mismo usuario. Cada línea se cuenta de manera individual.

   En este caso se trabajó a nivel de conversación, es decir, por cada

conversación se calculó lo siguiente:

- Número de interacciones generadas por el usuario acosador.
- Número de interacciones generadas por la víctima.
- Número de líneas generadas por el usuario acosador.
- Número de líneas generadas por la víctima.
- Usuario que inicia la conversación.

Las figuras 13, 14, 15 y 16, 17 presentan estadísticas generadas a partir de los datos anteriores. En las gráficas de las figuras 13 y 14 se puede observar que ambos usuarios (víctima y depredador), tienen en promedio la misma cantidad de interacciones. Por otra parte, con las figuras 15 y 16 se puede ver que el número promedio de líneas generadas para ambos usuarios, es ligeramente mayor para el caso del usuario acosador. Finalmente, en la figura 17 se puede visualizar que en la mayoría de las conversaciones del corpus, el usuario que inicia la conversación es el usuario acosador, con casi el doble de veces que la víctima.

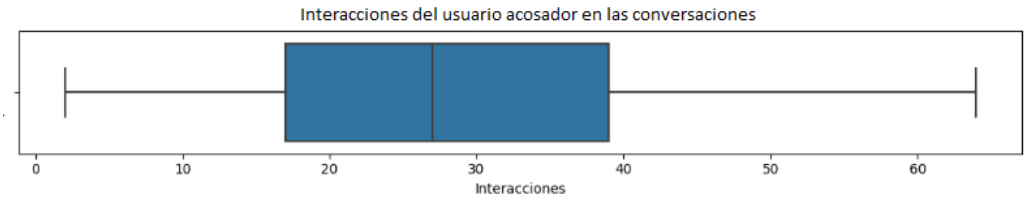

Figura 13: Cantidad de Interacciones generadas por el acosador en las Conversaciones. (Promedio: 28.44 | Desviación Estándar: 13.84)

4. **Tiempo de espera entre interacciones de los usuarios:**

   Estas estadísticas se realizaron con el objetivo de observar el tiempo que espera cada usuario para continuar la conversación. Para ello, las estadísticas se realizaron a nivel de conversación y tomando en cuenta una distribución de turnos por interacciones (recordemos que una interacción se refiere a todo el conjunto de líneas escritas por un usuario, hasta que el otro toma el turno). De esta forma, se captura

Figura 14: Cantidad de Interacciones generadas por la víctima en las Conversaciones. (Promedio: 28.32 | Desviación Estándar: 13.79)

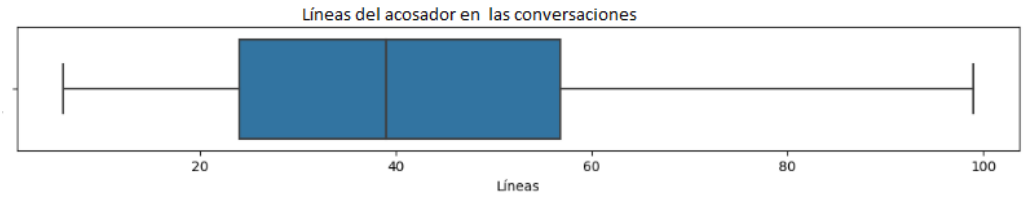

Figura 15: Cantidad de líneas generadas por el acosador en las Conversaciones. (Promedio: 40.63 | Desviación Estándar: 20.50)

Figura 16: Cantidad de líneas generadas por la víctima en las Conversaciones. (Promedio: 39.83 | Desviación Estándar: 18.95)

Figura 17: Cantidad de veces que cada usuario inicia una conversación. (Acosador: 932 | Víctima: 534)

el tiempo que espera un usuario en contestar, desde la última intervención del otro usuario. Los tiempos de espera se guardan en minutos, debido a que el formato que se maneja en el corpus es el siguiente: (HH:MM). Posteriormente, por cada conversación se calculó lo siguiente:

- El promedio de tiempo de espera del usuario acosador entre interacciones.
- El promedio de tiempo de espera de la víctima entre interacciones.

Las figuras 18 y 19 presentan gráficas de las estadísticas realizadas con los datos anteriores. En estas gráficas se puede observar que el usuario víctima responde ligeramente más inmediato que el usuario acosador. Esto, resulta interesante analizar, ya que, se puede intuir que el usuario acosador logra mantener la atención de la víctima.

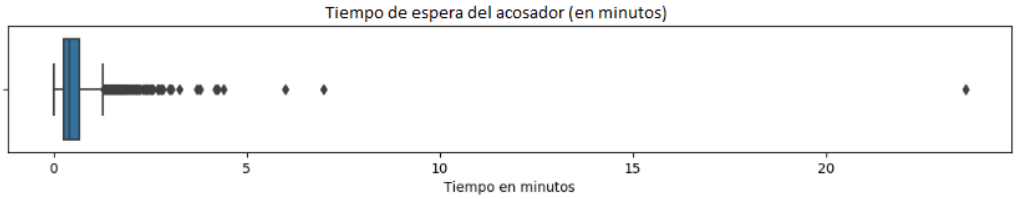

Figura 18: Tiempo de espera del acosador en las conversaciones. (Promedio: 0.57 | Desviación Estándar: 0.81)

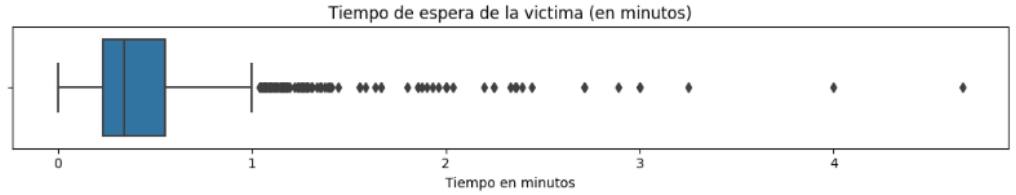

Figura 19: Tiempo de espera de la víctima en las conversaciones. (Promedio: 0.45 | Desviación Estándar: 0.40)

5. **Preguntas realizadas por conversación:**

El objetivo en estas estadísticas es analizar la cantidad de líneas que contienen preguntas realizadas por los usuarios, en una conversación. Para ello, en cada línea de conversación se buscan las siguientes estructuras:

- Preguntas (con signo): Si la línea de conversación contiene algún signo de interrogación "?".
- Wh Words + ?: Si la línea de conversación contiene alguna palabra considerada como wh words seguido de signo de interrogación. (Wh Words = what, where, when, who y why).

Dado lo anterior, por cada conversación se calculó la siguiente:

- Total de líneas con preguntas generadas por el usuario acosador.
- Total de líneas con preguntas generadas por la víctima.
- Total de líneas con Wh-words generadas por el usuario acosador.
- Total de líneas con Wh-words generadas por la víctima.

Las figuras 20, 21, 22, 23 presentan gráficas de las estadísticas realizadas con los datos anteriormente descritos. Con estas gráficas, se puede observar que la cantidad promedio de líneas generadas por los usuarios con preguntas y wh-words es ligeramente proporcional en ambos casos.

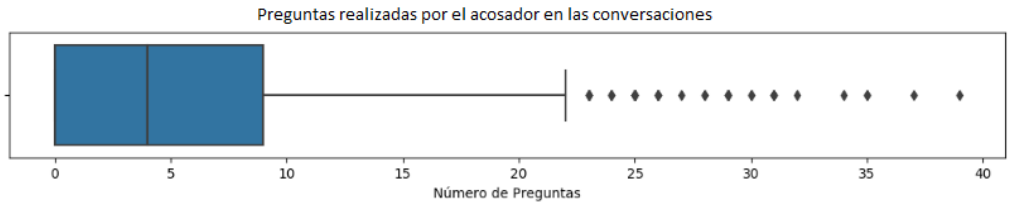

Figura 20: Cantidad de preguntas realizadas por el acosador en las conversaciones. (Promedio: 5.84 | Desviación Estándar: 6.77)

6. **Frecuencias relativas de categorías LIWC:**

   Estas estadísticas tienen como objetivo observar la presencia de cada categoría LIWC en una conversación en términos de porcentajes.

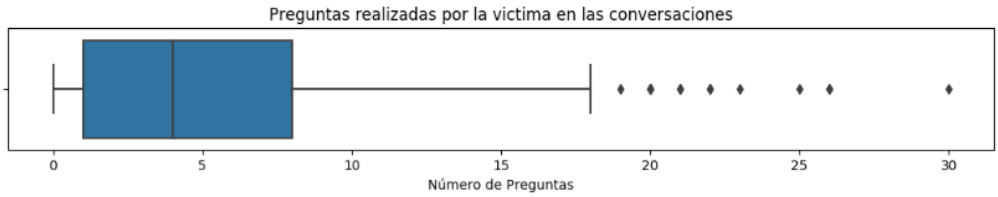

Figura 21: Cantidad de preguntas realizadas por la víctima en las conversaciones. (Promedio: 4.97 | Desviación Estándar: 4.51)

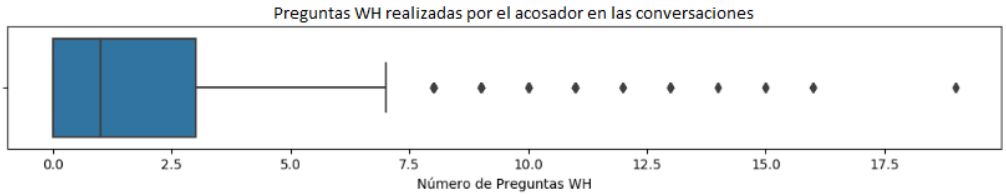

Figura 22: Cantidad de preguntas "WH" realizadas por el acosador en las conversaciones. (Promedio: 2.16 | Desviación Estándar: 2.94)

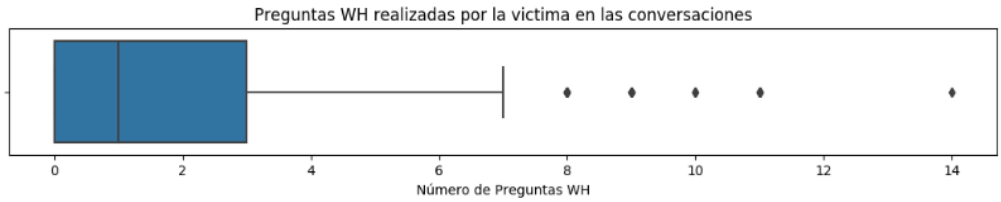

Figura 23: Cantidad de preguntas "WH" realizadas por la víctima en las conversaciones. (Promedio: 1.64 | Desviación Estándar: 2.08)

Para ello, por cada conversación se capturaron por separado todos los textos de cada usuario (acosador y víctima), teniendo de esta forma, en un archivo todas las conversaciones con solo los textos de los usuarios pedófilos y en otro archivo todas las conversaciones con los textos de las víctimas. Posteriormente, ambos archivos se procesaron de la siguiente manera:

- Todo el texto se transformó a minúsculas.
- Se eliminaron todos los signos de puntuación e interrogación.
- Se eliminaron los espacios extras.

- Se eliminaron los números.
- Se eliminaron los urls.

Una vez hecho esto, se calculó la frecuencia de palabras pertenecientes a cada categoría LIWC y posteriormente se procedió a calcular la presencia de cada una de estas en términos de porcentaje. Para ello, en cada texto, se calculó el porcentaje de frecuencia de cada categoría LIWC con respecto al total de palabras del texto. Por ejemplo, un texto con un total de 80 palabras y de estas 12 son de la categoría A, esta categoría tiene un 15 % de presencia en el texto.

Las figuras 24, 25, 26, 27, 28 presentan gráficas con las frecuencias relativas de las categorías LIWC presentes en ambos textos (acosador y víctima). Para una mejor visualización, las 68 categorías LIWC se dividieron en los siguientes 4 grupos de acuerdo con el sitio web oficial de LIWC[10]:

a) **Dimensiones lingüísticas:** Pronombres, Yo, Nosotros, Unomismo, Tu, Otro, Negaciones, Afirmaciones, Artículos, Preposiciones, Numeros.

b) **Procesos psicológicos:** Afectivo, Emociones positivas, Sentimientos positivos, Optimismo, Emociones negativas, Ansiedad, Enojo, Tristeza, Procesos cognitivos, Causa, Entendimiento, Discrepancia, Inhibición, Tentativo, Certeza, Sentidos, Ver, Oír, Sentir, Social, Comunicación, Referencia a otras personas, Amigos, Familia, Humanos.

c) **Relatividad:** Tiempo, Pasado, Presente, Futuro, Espacio, Arriba, Abajo, Inclusivo, Exclusivo, Moción.

d) **Asuntos personales:** Ocupación, Escuela, Trabajo, Logro, Placer, Casa, Deportes, TV, Música, Dinero, Metafísico, Religión, Muerte, Físico, Cuerpo, Sexual, Comer, Dormir, Asearse.

e) **Dimensiones experimentales:** Groserías, No fluido, Rellenos.

Los colores de las barras en las gráficas (19-23) están denotados por los siguientes colores: rojo para el usuario acosador y verde para la víctima.

[10] http://www.liwc.net/liwcespanol/descriptiontable1.php

Para el grupo 1 de categorías LIWC (Dimensiones lingüísticas), se tiene que la mayoría de las categorías se encuentran más presentes en los textos del usuario acosador. Algunas de las más presentes fueron: Pronombres, Yo y Preposiciones. Mientras que para la categoría Afirmaciones el usuario con más presencia fue la víctima.

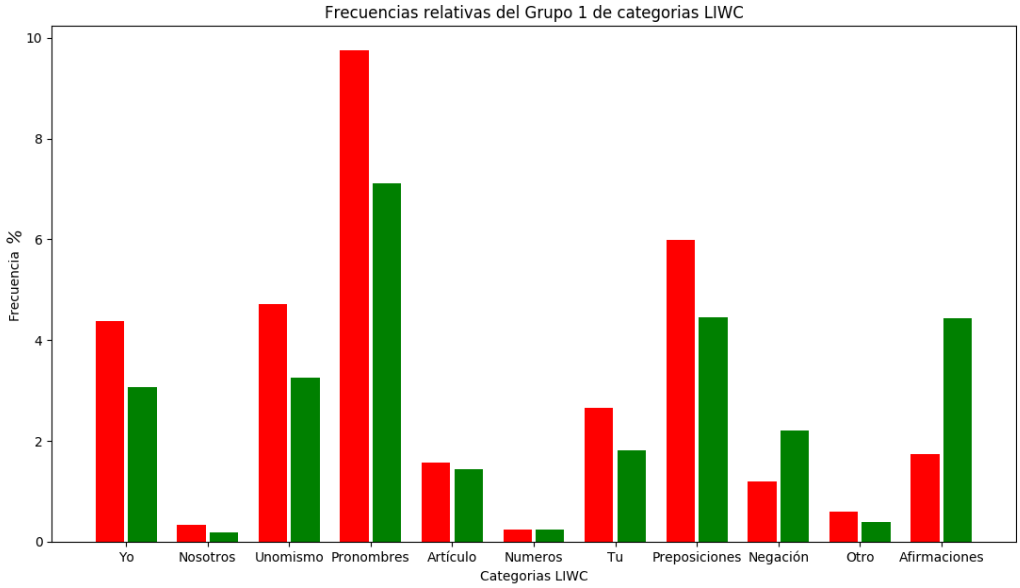

Figura 24: Frecuencias relativas del Grupo 1 de categorías LIWC presentes en ambos textos.

Para el grupo 2 de categorías LIWC (Procesos psicológicos), se tiene que las categorías más presentes en los textos del usuario pedófilo fueron: Social, Referencia a otras personas, Procesos cognitivos y Discrepancias. Por otra parte, las categorías más presentes en los textos de la víctima fueron: Emociones positivas, Afectivo y Causa. Nuevamente, se puede ver que la mayoría de las categorías resultan estar más presentes en los textos del usuario acosador.

En el caso del grupo 3 de categorías LIWC (Relatividad), se observa que la categoría más presente en ambos textos es la categoría: Presente. Mientras que las demás categorías se encuentran proporcionalmente presentes en ambos textos.

Para el grupo 4 de categorías LIWC (Asuntos personales), se tiene

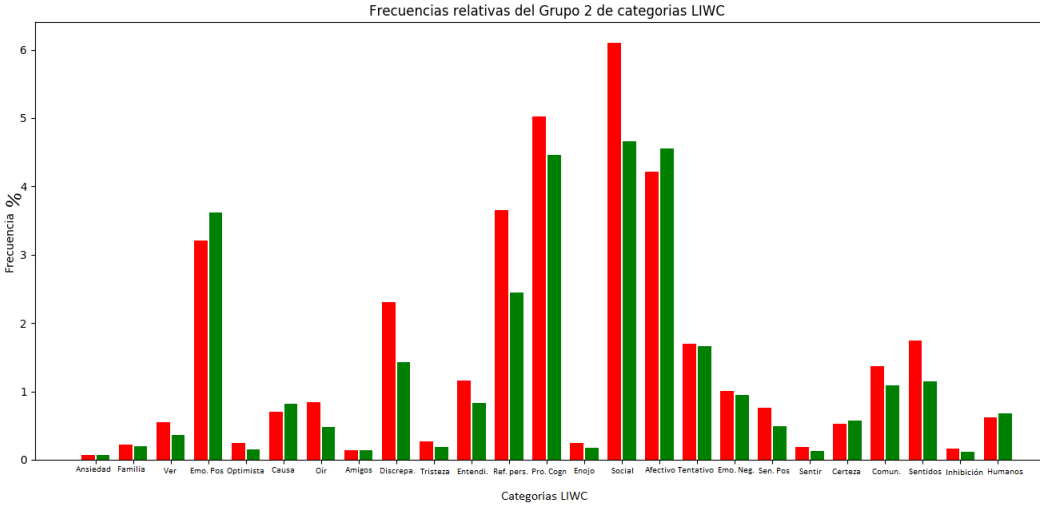

Figura 25: Frecuencias relativas del Grupo 2 de categorías LIWC presentes en ambos textos.

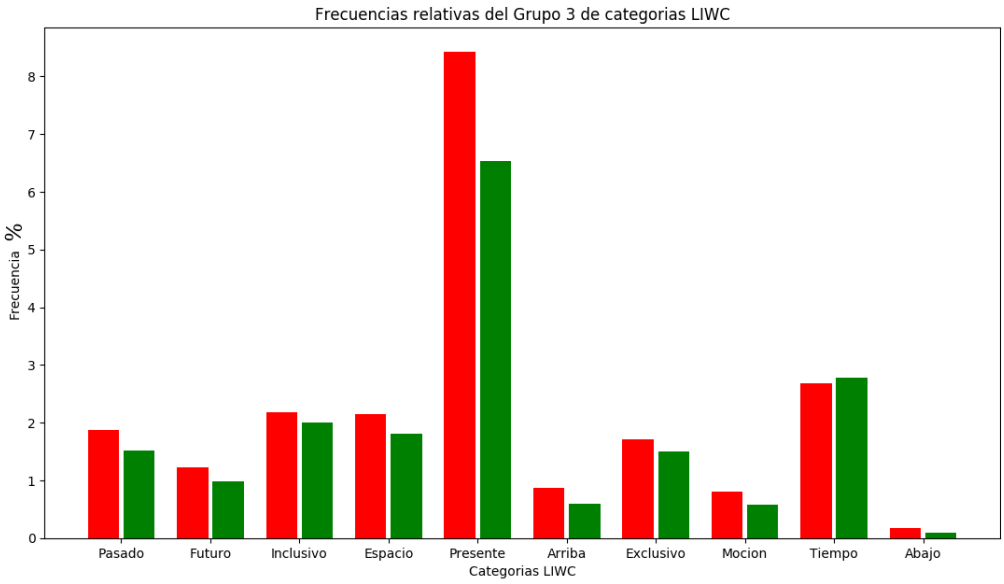

Figura 26: Frecuencias relativas del Grupo 3 de categorías LIWC presentes en ambos textos.

que ninguna de estas se encuentra mayormente en los textos de la víctima, es decir, todas estas están más presentes en los textos del usuario pedófilo.

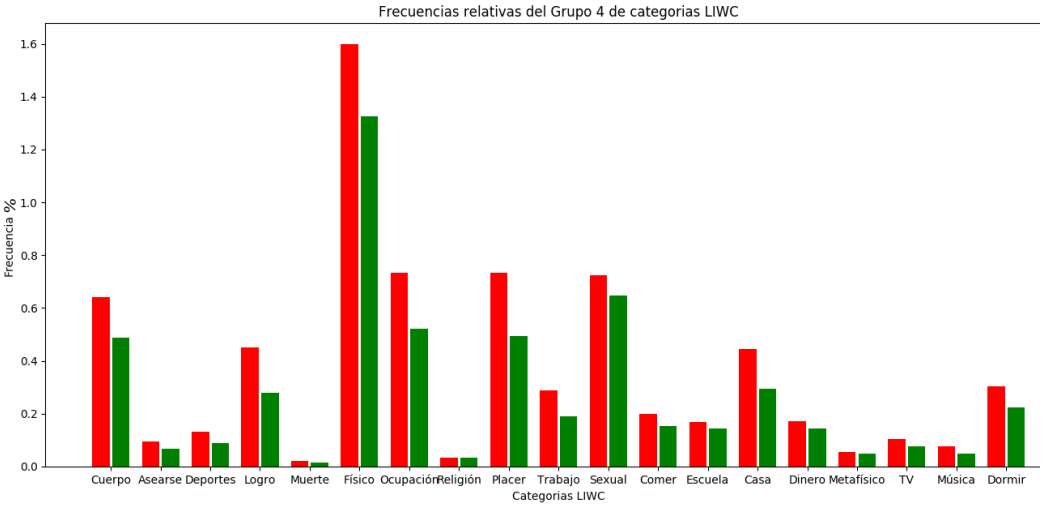

Figura 27: Frecuencias relativas del Grupo 4 de categorías LIWC presentes en ambos textos

Finalmente, para el grupo 5 de categorías LIWC (Dimensiones experimentales), las más presentes en los textos del acosador, fueron: Groserías y Rellenos.

Las estadísticas realizadas con el corpus de conversaciones, permite conocer datos interesantes de los usuarios presentes en la conversación (acosador y víctima); sin embargo, muchos de estos datos son importantes a nivel de conversación. En este caso, para este proyecto terminal interesa analizar características a nivel de línea. Es decir, se necesita un conjunto de datos compuesto por líneas de conversación del cual, se puedan representar sus atributos y utilizarlos para entrenar un modelo de predicción que ayude a identificar líneas de conversación que son más incriminatorias de acoso sexual. Es por esta razón, que se decidió procesar el corpus de conversaciones para guardar únicamente las líneas de conversación escritas por el usuario pedófilo. Este procesamiento permitió obtener un conjunto de datos compuesto por líneas de conversación pedófilas y no por conversaciones enteras con ambos usuarios.

Consecuentemente, cada línea de conversación almacenada se eti-

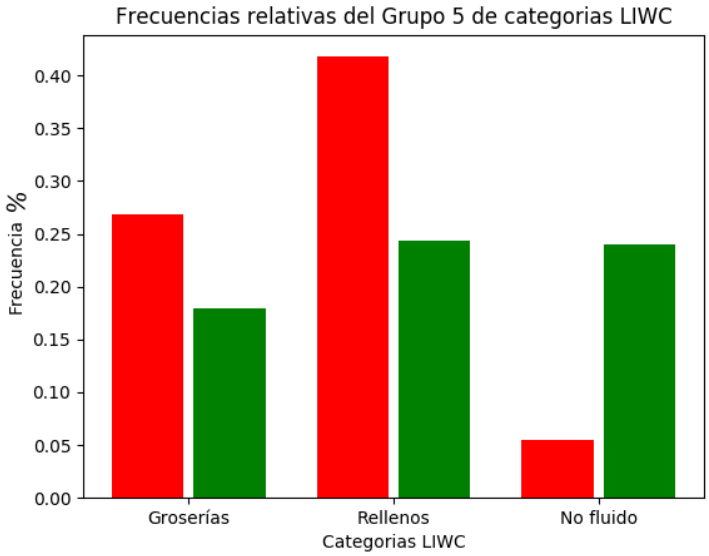

Figura 28: Frecuencias relativas del Grupo 5 de categorías LIWC presentes en ambos textos

queto como Incriminatoria o No Incriminatoria, de acuerdo al archivo de solución: *Gold Standard* para la tarea 2 (Identificación de líneas pedófilas) en la competencia CLEF PAN 2012. De esta forma, se tiene un conjunto de datos etiquetado que sera denominado: corpus de intervenciones.

Finalmente, la figura 29 muestra una gráfica con la distribución de ambas clases (Incriminatoria o No Incriminatoria), así como también el número total de líneas contenidas en el corpus de intervenciones. De esta figura se puede observar que se cuenta con un corpus bastante desbalanceado.

4.2. Atributos

En este trabajo se definieron 4 diferentes familias de atributos para representar los textos (líneas de conversación), los cuales se describen a continuación:

- Atributos Textuales: Estos atributos tienen como objetivo de captu-

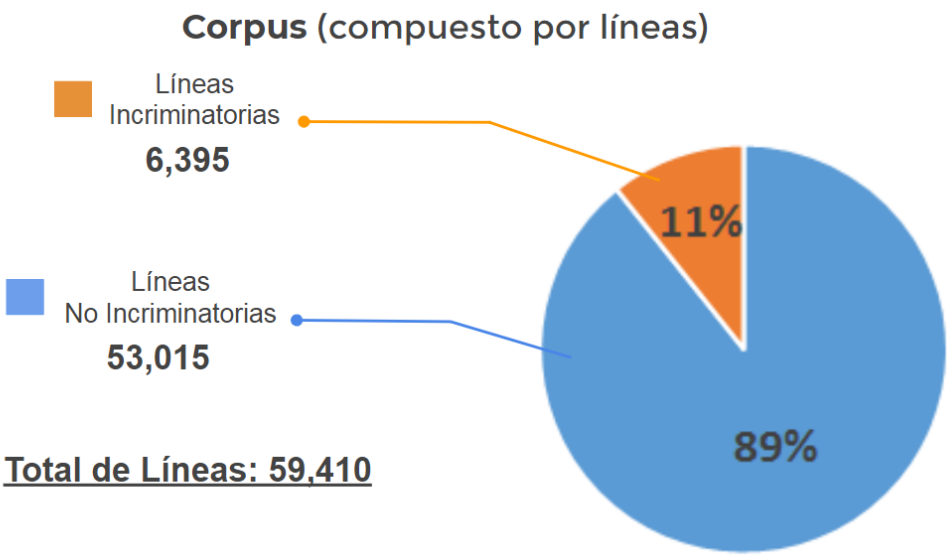

Figura 29: Distribución de clases contenidas en el corpus de intervenciones.

rar todas aquellas palabras que se encuentran presentes en una línea de conversación.

- Atributos de Estilo: Tienen el objetivo de capturar el estilo propio del autor a través de la captura de características como la cantidad de palabras, de números, de signos de puntuación y de preguntas que se encuentran presentes en las líneas de conversación.
- Atributos Contextuales (LIWC): Estos atributos tienen como objetivo capturar aquellas categorías de palabras que se encuentran presentes en las líneas de conversación, principalmente los temas de conversación de los que se habla. Para la captura de estos atributos, se hizo uso de LIWC[11], un recurso léxico que permite analizar textos. El cual, clasifica cada palabra en una de las diferentes categorías de palabras que maneja LIWC.
- Atributos Gramaticales (POS): Tienen el objetivo de capturar ciertas construcciones gramaticales, que son utilizadas en las líneas de con-

[11] http://liwc.wpengine.com/

versación. Para esto, se realizó un proceso de etiquetado POS (*Part-Of-Speech*), el cual es el proceso de asignar o etiquetar a cada una de las palabras de un texto su categoría gramatical. Las categorías gramaticales que se utilizaron son las 36 etiquetas POS contenidas en el etiquetador TreeTagger[12].

La tabla 2, resume cada uno de los atributos descritos anteriormente, así como un ejemplo de estos.

Atributo	Descripción	Ejemplo con: *i hope you like me when we meet*
Textuales	Información textual referente a cada línea de conversación.	`i hope you like me when we meet` (todas las palabras)
Estilísticos	Cantidad de palabras, números, preguntas y signos.	`Palabras` - 8, `Números` - 0, `Sig. puntuación` - 0, `Preguntas` - 0
Contextuales	Categorías LIWC que se encuentran presentes en las líneas de conversación.	`hope` - Afectiva, Optimista; `like` - Afectiva, Presente; `meet` - Social, Presente
Gramaticales	Construccion de etiquetas POS generadas a partir de cada línea de conversación.	`i` - NNS, `hope` - VBP, `you` - PP, `like` - VBP, `me` - PP, `when` - WRB, `we` - PP, `meet` - VBP

Tabla 2: Descripción de los atributos.

4.2.1. Representación

Para la representación de los atributos textuales, se utilizó una Bolsa de Palabras (BoW), con los 10K términos más frecuentes en el corpus.

Los atributos de estilo representan conteos absolutos de la presencia de alguna de las características, por lo tanto, una vez que estos atributos son contabilizados, se hace un proceso de normalización con la finalidad de llevarlos a una escala entre 0 y 1.

[12]https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

En cuanto a los atributos contextuales, su representación se basó en el porcentaje de presencia de cada categoría LIWC por cada una de las líneas de conversación. Es decir, por cada línea se calculaba la presencia de cada categoría LIWC en términos de porcentajes.

Por último, para los atributos gramaticales, se utilizó nuevamente una bolsa de palabras (BoW) de las 10K etiquetas POS más frecuentes.

4.3. Experimentos

A continuación se describen, cada uno de los experimentos desarrollados en este proyecto terminal. Estos experimentos se realizaron, utilizando el corpus de intervenciones, y con el objetivo de encontrar un método que permitiera identificar líneas incriminarías de acoso sexual.

Se realizaron dos grandes grupos de experimentos, el primero utilizando cada tipo de atributo de manera individual, mientras que el segundo grupo de experimentos incorporo combinaciones de los 4 diferentes tipos de atributos.

Para cada experimento se probaron 5 diferentes algoritmos de aprendizaje: Naïve Bayes, SVM, K-Nearest Neighbors, Decision Tree y Random Forest, cada uno de estos métodos se obtuvieron de la biblioteca Scikit-learn[13]. Mientras que, para la evaluación, se realizó una validación cruzada de 10 pliegues y se utilizó la métrica F1 de la clase positiva (clase Incriminatoria). Adicionalmente, para cada experimento con representación en bolsa de palabras (BoW), se utilizaron 4 distintas configuraciones:

- Configuración 1: Palabras con pesado booleano.
- Configuración 2: Bi-gramas de palabras y pesado booleano.
- Configuración 3: Tri-gramas de palabras y pesado booleano.
- Configuración 4: Palabras con pesado Tf.

[13]https://scikit-learn.org/stable/

Debido a que al ser atributos con características textuales los n-gramas de diferentes tamaños permiten capturar información valiosa que puede estar contenida en los textos.

4.3.1. Atributos Individuales

Este primer grupo de experimentos, se realizaron con el objetivo principal de identificar que tan valioso es cada tipo de atributo de forma individual. Es decir, que tanta información aporta a un modelo de aprendizaje para que este pueda identificar una línea incriminatoria.

La tabla 3 enumera cada experimento realizado en este primer grupo de experimentos.

Experimento	Atributos	Abreviatura
1	Textuales	T
2	Estilo	E
3	Contextuales	L
4	Gramaticales	P

Tabla 3: Experimentos realizados utilizando atributos individuales.

A continuación se describen los experimentos realizados y sus resultados.

- Experimento 1 - Atributos textuales (T): En este experimento se utilizaron solo los atributos textuales. En la tabla 4 se muestran los resultados obtenidos en este experimento, en el cual los mejores resultados se obtuvieron con el algoritmo SVM con un F1-Score de **0.59** con pesado booleano y tf en unigramas. Seguido del algoritmo Random Forest con un F1-Score de **0.48** también para los unigramas en pesado booleano y tf.
- Experimento 2 - Atributos de estilo (E): En este experimento se utilizaron solo atributos de estilo. Para este experimento los mejores resultados fueron obtenidos con el algoritmo Naïve Bayes con un F1-Score de **0.40** y Random Forest con un F1-Score de **0.41** utilizando una representación normalizada en escala de 0 a 1.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.50	**0.59**	0.38	0.23	0.48
Rep.2: 2-gramas (bool)	0.44	0.45	0.35	0.13	0.43
Rep.3: 3-gramas (bool)	0.26	0.26	0.29	0.05	0.30
Rep.4: 1-gramas (Tf)	0.50	**0.59**	0.38	0.23	0.48

Tabla 4: Resultados obtenidos del experimento 1.

- Experimento 3 - Atributos contextuales (L): En este experimento se utilizaron solo los atributos Contextuales. El mejor resultado de este experimento alcanzo un F1-Score de **0.46** utilizando una representación normalizada en escala de 0 a 1.
- Experimento 4 - Atributos gramaticales (P): En este experimento se utilizaron solo los atributos gramaticales. En la tabla 5 se muestran los resultados obtenidos en este experimento, en donde los mejores resultados fueron obtenidos con el algoritmo Naïve Bayes con un F1-Score de **0.36** para bigramas y **0.35** para trigramas.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.31	0.04	0.27	0.24	0.19
Rep.2: 2-gramas (bool)	**0.36**	0.03	0.28	0.30	0.25
Rep.3: 3-gramas (bool)	**0.35**	0.18	0.27	0.29	0.24
Rep.4: 1-gramas (Tf)	0.31	0.03	0.28	0.27	0.24

Tabla 5: Resultados obtenidos del experimento 4.

Con este primer grupo de experimentos, se puede observar que el atributo que mejor ayuda a la construcción de un modelo de predicción es el atributo: Textual. Los resultados de este atributo son los resultados más altos obtenidos hasta este momento con un F1-Score de **0.59**. Enseguida,

se tiene que el atributo Contextual, alcanza el segundo mejor resultado de este grupo de experimentos, con un F1-Score de **0.46**. Por otra parte, los atributos de estilo y POS no alcanzan buenos resultados como se esperaba. En el caso de las etiquetas POS, es probable que no sean muy precisas debido al tipo de escritura que normalmente se manejan en las conversaciones de chat. Estos resultados obtenidos se esperan mejorar con la combinación de atributos, los cuales se detallan en el siguiente apartado.

4.3.2. Atributos Combinados

Para este segundo grupo de experimentos, se realizó una segunda división de experimentos:

- Combinaciones con el atributo Textual: Esto con el objetivo de encontrar una combinación de atributos que ayudara a mejorar el resultado obtenido con el experimento 1.
- Combinaciones sin el atributo Textual: Esto con el objetivo de verificar si existe una combinación de los atributos de Estilo, Contextual y Gramatical que mejore los resultados obtenidos con el experimento 1.

En cuanto a los experimentos realizados utilizando combinaciones con el atributo textual, la tabla 6 enumera los experimentos realizados, y posteriormente se describe los resultados obtenidos.

Experimento	Atributos	Abreviatura
5	Textuales y de Estilo	T,E
6	Textuales y Contextuales	T,L
7	Textuales y Gramaticales	T,P
8	Textuales, Estilo, Contextuales y Gramaticales	T,E,L,P

Tabla 6: Experimentos realizados utilizando combinaciones con el atributo textual.

- Experimento 5 - Atributos textuales y de estilo (T,E): En este experimento se utilizaron los atributos textuales combinados con atributos

de estilo. En la tabla 7 se muestran los resultados obtenidos en este experimento, en donde los mejores resultados fueron obtenidos nuevamente con el algoritmo SVM con un F1-Score de **0.59** para los unigramas en pesado booleano y tf. Asimismo, los segundos mejores resultados fueron obtenidos con el algritmo Random Forest con un F1-Score de **0.48** también para los unigramas.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.50	**0.59**	0.40	0.51	0.48
Rep.2: 2-gramas (bool)	0.45	0.46	0.34	0.36	0.40
Rep.3: 3-gramas (bool)	0.26	0.26	0.31	0.25	0.26
Rep.4: 1-gramas (Tf)	0.50	**0.59**	0.40	0.50	0.48

Tabla 7: Resultados obtenidos del experimento 5.

- Experimento 6 - Atributos textuales y contextuales (T,L): En este experimento se utilizaron solo los atributos de tipo textual y contextual. En la tabla 8 se muestran los resultados obtenidos en este experimento, en donde los mejores resultados igualmente fueron obtenidos con el algoritmo SVM con un F1-Score de **0.59**.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.46	**0.59**	0.41	0.49	0.48
Rep.2: 2-gramas (bool)	0.44	0.47	0.38	0.46	0.43
Rep.3: 3-gramas (bool)	0.42	0.32	0.39	0.43	0.43
Rep.4: 1-gramas (Tf)	0.46	**0.59**	0.41	0.49	0.47

Tabla 8: Resultados obtenidos del experimento 6.

- Experimento 7 - Atributos textuales y gramaticales (T,P): En este experimento se utilizaron solo los atributos de tipo textual y gramati-

cal. En la tabla 9 se muestran los resultados obtenidos en este experimento, en donde los mejores resultados fueron obtenidos una vez más con el algoritmo SVM con un F1-Score de **0.59**.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.47	**0.59**	0.39	0.50	0.42
Rep.2: 2-gramas (bool)	0.43	0.46	0.30	0.37	0.30
Rep.3: 3-gramas (bool)	0.37	0.33	0.28	0.30	0.26
Rep.4: 1-gramas (Tf)	0.47	**0.59**	0.39	0.50	0.43

Tabla 9: Resultados obtenidos del experimento 7.

- Experimento 8 - Atributos textuales, de estilo, contextuales y gramaticales (T,E,L,P): En este experimento se combinaron los 4 tipos de atributos. En la tabla 10 se muestran los resultados obtenidos en este experimento, en donde los mejores resultados fueron obtenidos nuevamente con el algoritmo SVM con un F1-Score de **0.59**.

Algoritmos de aprendizaje	NB	SVM	K-N	DT	RF
	F1-Score	F1-Score	F1-Score	F1-Score	F1-Score
Rep.1: 1-gramas (bool)	0.43	**0.59**	0.41	0.49	0.46
Rep.2: 2-gramas (bool)	0.42	0.48	0.33	0.46	0.42
Rep.3: 3-gramas (bool)	0.41	0.41	0.31	0.43	0.42
Rep.4: 1-gramas (Tf)	0.43	**0.59**	0.41	0.50	0.46

Tabla 10: Resultados obtenidos del experimento 8.

Con estas combinaciones realizadas al atributo Textual, se obtiene nuevamente como resultado más alto un F1-Score de **0.59**. Esto nos deja claro que el atributo más presente es el atributo textual, ya que los resultados obtenidos en estos experimentos y el experimento 1 son muy similares.

También, se puede verificar que los demás atributos (Estilo Contextual y Gramatical) combinados con el atributo Textual, no ayudan a mejorar el F1-Score del experimento 1 como se esperaba, sin embargo, tampoco lo empeoran.

A continuación, se presentan los resultados obtenidos en los experimentos realizados con las combinaciones de atributos de Estilo, Contextual y Gramatical. La tabla 11 enumera los experimentos desarrollados con estas combinaciones.

Experimento	Atributos	Abreviatura
9	Estilo y Contextuales	E,L
10	Estilo y Gramaticales	E,P
11	Estilo, Contextuales y Gramaticales	E,L,P
12	Contextuales y Gramaticales	L,P

Tabla 11: Experimentos realizados utilizando combinaciones de los atributos de estilo, contextual y gramatical.

- Experimento 9 - Atributos de Estilo y Contextuales (E,L): En este experimento se utilizaron atributos de estilo combinados con atributos contextuales. Los mejores resultados de este experimento fueron obtenidos con los algoritmos K-Neighbors con un F1-Score de **0.45** y Random Forest con un F1-Score de **0.46** utilizando una representación normalizada en escala de 0 a 1 para ambos atributos.

- Experimento 10 - Atributos de Estilo y Gramaticales (E,P): En este experimento se utilizó una combinación de los atributos estilográficos y gramaticales. El mejor resultado de este experimento fue obtenido con el algoritmo Naïve Bayes con un F1-Score de **0.36** y una representación de bi-gramas con pesado booleano.

- Experimento 11 - Atributos de estilo, contextuales y gramaticales (E,L,P): En este experimento se combinaron tres tipos de atributos: de contexto, gramática y estilo. El mejor resultado de este experimento fue obtenido con el algoritmo Random Forest con un F1-Score de **0.47** y una representación de uni-gramas con pesado booleano.

- Experimento 12 - Atributos contextuales y gramaticales (L,P): En este experimento se utilizó una combinación de los atributos de contexto y gramática. El mejor resultado de este experimento fue obtenido con el algoritmo Random Forest con un F1-Score de **0.47** y una representación de uni-gramas con pesado tf.

Como resultados de estos últimos experimentos se tiene que ninguna de las combinaciones realizadas de los atributos de Estilo, Contextual y Gramatical superan el F1-Score de **0.59**, obtenido en los experimentos anteriores. Con esto último, se reafirma que el atributo con mejor aporte de información sigue siendo el atributo Textual. Si bien es notable que estos atributos también aportan cierto grado de información al modelo, esto no permite que el modelo de predicción tenga un grado de confiabilidad más alta.

4.3.3. Discusión

Como se puede observar en los resultados de los experimentos realizados, el F1-Score más alto obtenido fue de **0.59** en los experimentos 1, 5, 6, 7 y 8. Como se ha venido observando, este resultado fue obtenido en todos los casos con el algoritmo SVM, el cual al igual que Random Forest fueron los que se mantuvieron más estables en todos los experimentos desarrollados.

Analizando los resultados obtenidos se puede concluir que el atributo Textual es el que más aporta información en el aprendizaje de un modelo de predicción. Por el lado contrario, los atributos de estilo, contextuales y gramaticales, alcanzan un F1-Score de 0.59 cuando se combinan con el atributo textual; sin embargo, resultan de un score más bajo cuando se utilizan de manera individual o se combinan solo entre ellos.

Finalmente, se seleccionó como mejor modelo de predicción el obtenido en el experimento 6, ya que al analizar la matriz de confusión, este resulto tener menores valores para los falsos negativos y falsos positivos (estos conceptos fueron explicados en la sección 2), en comparación con los experimentos 1, 5, 7 y 8.

Esto último lleva a concluir que el modelo generado del experimento

6 (atributos de Texto y contextuales) ayuda significativamente a identificar de mejor manera la clase positiva (líneas incriminatorias) del problema de clasificación que se aborda en la solución a la problemática planteada en este proyecto terminal. La figura 30 muestra la matriz de confusión obtenida con el experimento 6. Mientras que la figura 31 muestra la matriz de confusión del experimento 8, el segundo mejor resultado.

		Predicción	
	F1-Score: 0.59	**Positivo**	**Negativo**
Verdaderos	**Positivo**	3107	3288
	Negativo	993	52022
	Precisión	0.76	
	Recall	0.49	

Figura 30: Matriz de confusión del experimento 6 (Textual y Contextual).

		Predicción	
	F1-Score: 0.59	**Positivo**	**Negativo**
Verdaderos	**Positivo**	3118	3277
	Negativo	1022	51993
	Precisión	0.75	
	Recall	0.48	

Figura 31: Matriz de confusión del experimento 8 (Textual, Estilo, Contextual y Gramatical).

En la matriz de confusión del experimento 6 se puede observar que los valores de precisión y *recall* son ligeramente mayor a los valores de la matriz de confusión del experimento 8 (el segundo mejor resultado).

Adicional a lo anterior, se compara el mejor resultado con el obtenido en CLEF PAN 2012, en el cual se utilizó una métrica de evaluación $F_{\beta=3}$. En este caso, utilizando un algoritmo SVM y una representación de BoW de uni-gramas alcanzamos un $F_{\beta=3}$ de 0.4979, logrando superar el mejor resultado reportado en CLEF PAN 2012, en donde su $F_{\beta=3}$ es de 0.4762.

5. Desarrollo del Sistema Web

En esta sección se describe el funcionamiento del sistema web desarrollado para el apoyo en la identificación de líneas incriminatorias en una conversación de acoso sexual, así como también, se detalla la estructura de los módulos que lo componen. Este sistema web se encuentra compuesto de cuatro módulos principales, los cuales son: Modulo de carga, Modulo de procesamiento, Modulo de predicción y Modulo de Visualización.

A continuación, se muestra el esquema general del sistema, asimismo se describen cada uno de los módulos que lo integran.

5.1. Esquema general del sistema web

En la figura 32 se puede apreciar el esquema general del sistema web, así como cada uno de los módulos que lo componen. Cada módulo se describe a detalle en las siguientes subsecciones.

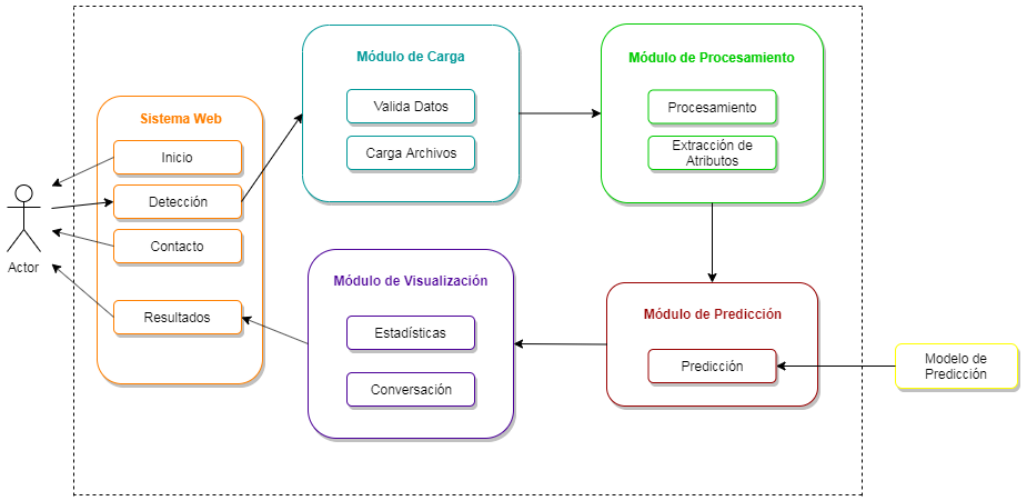

Figura 32: Esquema general del sistema web.

Como se puede observar en la figura 32, el recuadro naranja muestra las funciones principales del sistema web con las cuales interactúa el usuario, estas son: página de inicio, detección, contacto y resultados. La

página de detección se encarga de desencadenar los procesos correspondientes para analizar una conversación de acoso sexual, pasando por cada uno de los módulos contenidos para finalmente mostrar una visualización en la página de resultados. En las siguientes subsecciones se describen cada uno de los módulos que conforman el sistema web.

5.1.1. Módulo de carga

Este módulo recibe como entrada un archivo en formato XML el cual contiene una conversación de acoso sexual ya identificada, así como también el ID del usuario identificado como acosador. Una vez recibidos estos datos, el módulo se encarga de validar que estos sean archivos válidos. Para el caso del archivo que contiene la conversación de acoso, se valida que este sea un archivo xml. Para el ID del usuario acosador, se verifica que este ID realmente exista dentro de la conversación.

Posteriormente, una vez validados los datos correctamente se procede a guardar la conversación en un directorio para ser utilizado por los siguientes módulos. En la figura 33, se puede observar el proceso que realiza este módulo.

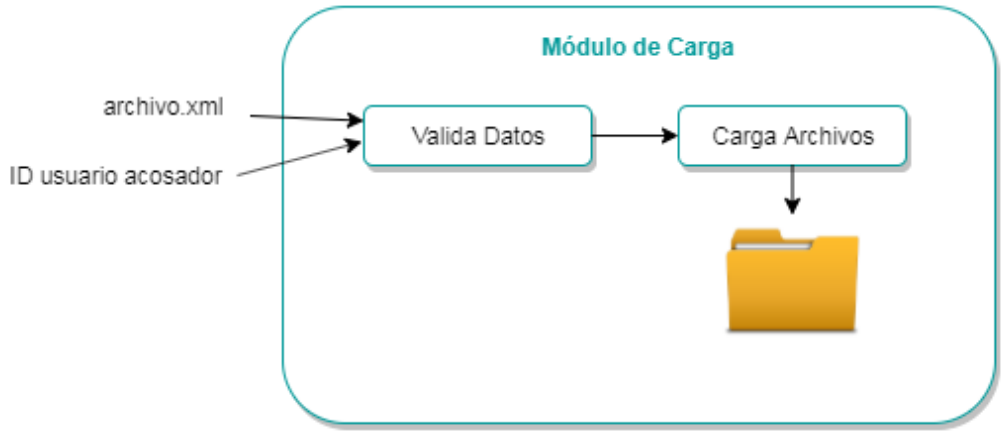

Figura 33: Módulo de carga del sistema web.

5.1.2. Módulo de procesamiento

En este módulo, se procede a realizar el procesamiento del archivo de conversación almacenado. Primero, se recupera el archivo de conversación, posteriormente, se realiza el procesamiento en donde se generan 5 archivos más, los cuales son:

- C+ID.txt: Este archivo de texto almacena toda la conversación por líneas. Es decir, cada línea del archivo de texto corresponde a una línea de conversación, en donde se almacena únicamente el ID de la línea, el usuario (acosador/víctima), la hora y el texto escrito.
- L+ID.txt: Este archivo almacena todas aquellas líneas de la conversación escritas únicamente por el usuario acosador.
- LV+ID.txt: Este archivo almacena todas aquellas líneas de la conversación escritas únicamente por el usuario víctima.
- F+ID.csv: Este archivo separado por comas, guarda el porcentaje de cada categoría LIWC presente en las líneas escritas por el usuario acosador.
- FV+ID.csv: Este archivo separado por comas, guarda el porcentaje de cada categoría LIWC presente en las líneas escritas por el usuario víctima.
- P+ID.txt: Este archivo de texto, almacena las etiquetas gramaticales de cada línea de conversación escrita por el usuario acosador.

Cada archivo que se genera aporta información específica que será utilizada por distintos métodos en los siguientes módulos.

Finalmente, una vez generados los archivos necesarios, se procede a realizar la extracción de los atributos contenidos en las líneas de conversación, que ayudarán a identificar líneas incriminatorias de acoso sexual. Para esto, se utilizan los archivos L+ID.txt, F+ID.csv y P+ID.txt, se extraen los atributos y se genera un nuevo archivo csv que contiene los atributos de cada línea de conversación escrita por el usuario acosador: G+Id.csv. La figura 34, muestra el proceso que realiza el módulo de procesamiento.

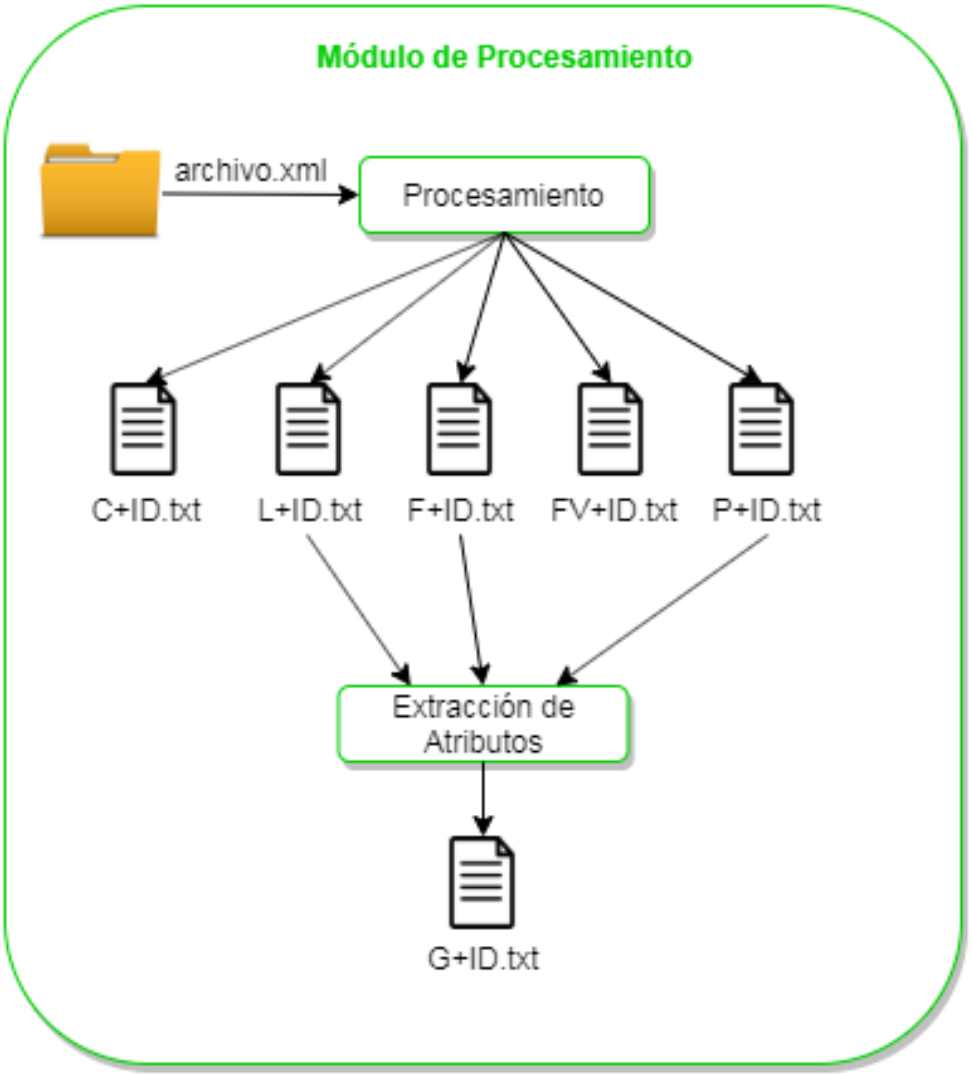

Figura 34: Módulo de procesamiento del sistema web.

5.1.3. Módulo de predicción

Este módulo se encarga de realizar la identificación de las líneas incriminatorias presentes en la conversación. Para ello, se recupera el archivo G+ID.txt, y posteriormente, se hace uso del archivo objeto que contiene el modelo de predicción generado anteriormente (sección 4.3), el cual devuelve como salida el ID de las líneas identificadas como incriminatorias. Estos ID son almacenados en un archivo de texto: D+ID.txt, para ser utilizado en el siguiente modulo. La figura 35, muestra el procedimiento que realiza este módulo.

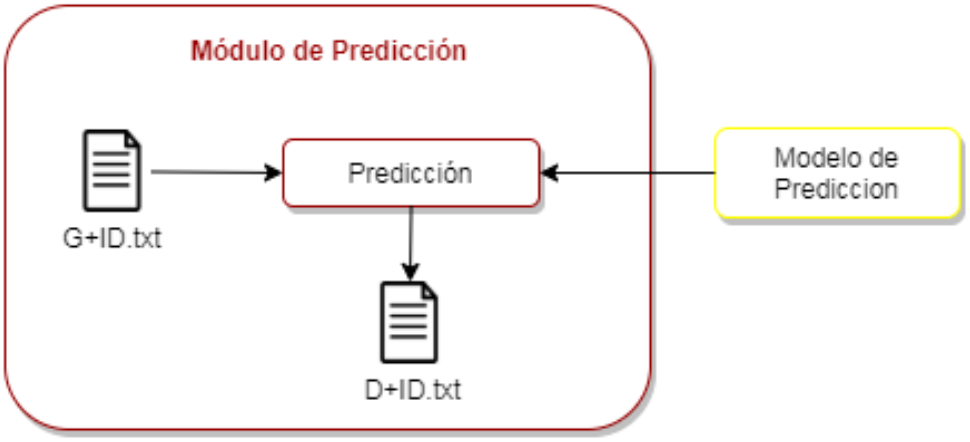

Figura 35: Módulo de predicción del sistema web.

5.1.4. Módulo de visualización

En este módulo, se cargan y procesan los datos necesarios para la visualización correspondiente. Para la visualización de estadísticas, se recuperan los archivos C+ID.txt, F+ID.csv, FV+ID.csv y D+ID.txt, a partir de los cuales se calculan los siguientes datos:

- Número de líneas del usuario acosador, víctima y totales.
- Rango de horario en que se desarrolla la conversación.

- Usuario que inicia la conversación.
- Distribución de toma de turno entre el usuario acosador y víctima.
- Número total de palabras, preguntas, signos y números de ambos usuarios.
- Porcentaje de presencia de cada categoría LIWC de las líneas del acosador.
- Porcentaje de presencia de cada categoría LIWC de las líneas de la víctima.
- Total de líneas identificadas como incriminatorias.

Por otra parte, para la visualización de la conversación, se recuperan los archivos, C+ID.txt y D+ID.txt, los cuales serán utilizados para representar los textos de cada línea de conversación y si esta es o no identificada como incriminatoria. La figura 36, proporciona un esquema del procedimiento que realiza este módulo.

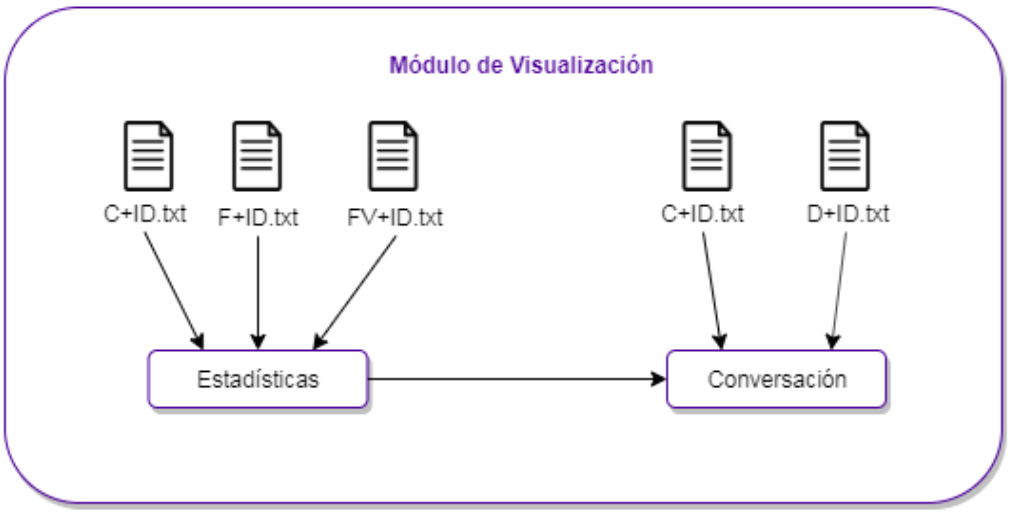

Figura 36: Módulo de visualización del sistema web.

5.2. Vistas del sistema web

En las figuras 37, 38, 39, 40 y 41 se pueden apreciar algunas páginas que componen el sistema web.

En la figura 37 se presenta la página de inicio del sistema web. En está, se muestra breve información sobre el Grooming, así como también un botón “Analizar conversación” que lleva a la página de detección. Asimismo, cuenta con un menú de navegación con enlaces a las páginas de contacto, detección e inicio.

Figura 37: Página de inicio del sistema web.

La figura 38 presenta la página de detecta del sistema web. En esta página se proporciona un breve formulario, donde se carga el archivo de conversación en formato XML, y se captura el ID del usuario acosador dentro de la conversación. También, se proporciona un enlace a un ejemplo de la estructura xml solicitada para el archivo de conversación.

En la figura 39 se puede visualizar la página de análisis del sistema web. Esta página nos notifica cuando la carga de archivos fue exitosa y proporciona un botón para continuar con el análisis de los datos. Una vez presionado el botón de continuar, aparece una barra de retroalimentación, donde se puede ver que el sistema se encuentra analizando los datos

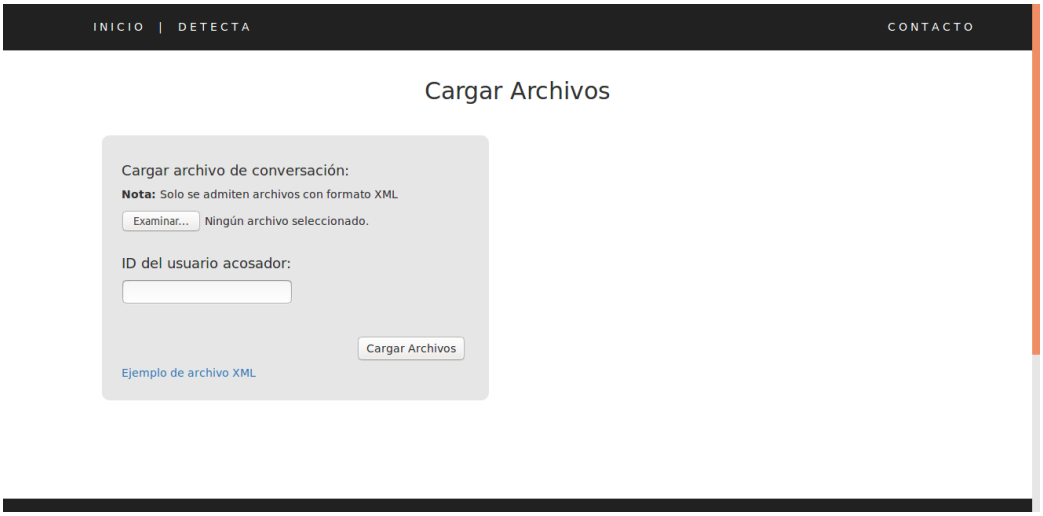

Figura 38: Página de detecta del sistema web.

proporcionados.

La figura 40 muestra la página de estadísticas del sistema web. En esta página se carga la visualización correspondiente a cada una de las estadísticas calculadas. Las primeras gráficas corresponden al número total de líneas escritas por ambos usuarios y el rango de horario en que se desarrolla la conversación. Consecutivamente, las siguientes visualizaciones muestran el usuario que inicia la conversación y la distribución de toma de turnos entre los usuarios, así como el promedio del tiempo que espera cada usuario en contestar. También muestra información relacionada al número de palabras, números, signos de puntuación y preguntas que realiza cada usuario dentro de la conversación. Finalmente, las ultimas gráficas muestran información acerca de los grupos de categorías LIWC más presentes para cada usuario de la conversación. En la parte inferior de la página, se encuentra un botón que lleva a la siguiente visualización, la cual corresponde a la conversación proporcionada y la identificación de las líneas etiquetadas como incriminatorias por el modelo de predicción.

Con la figura 41, se puede visualizar la página de conversación del sistema web. En esta página se carga la visualización correspondiente a la conversación proporcionada. Por el lado derecho se tienen las líneas escritas por el usuario acosador mientras que por el lado izquierdo se encuen-

Figura 39: Página de análisis del sistema web.

tras las líneas escritas por la víctima, Asimismo, se encuentran marcadas de color rojo aquellas líneas de conversación que el modelo clasifico como incriminatorias.

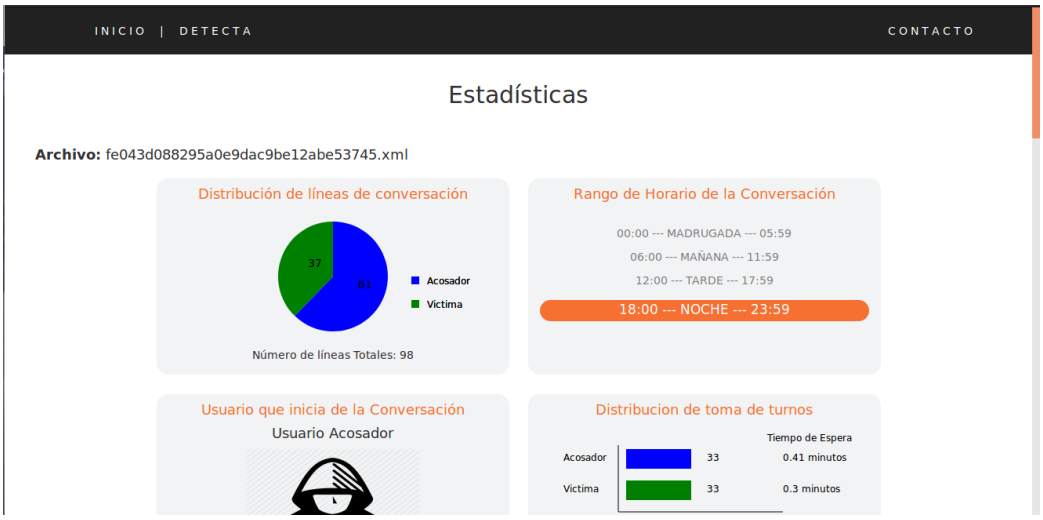

Figura 40: Página de estadísticas del sistema web.

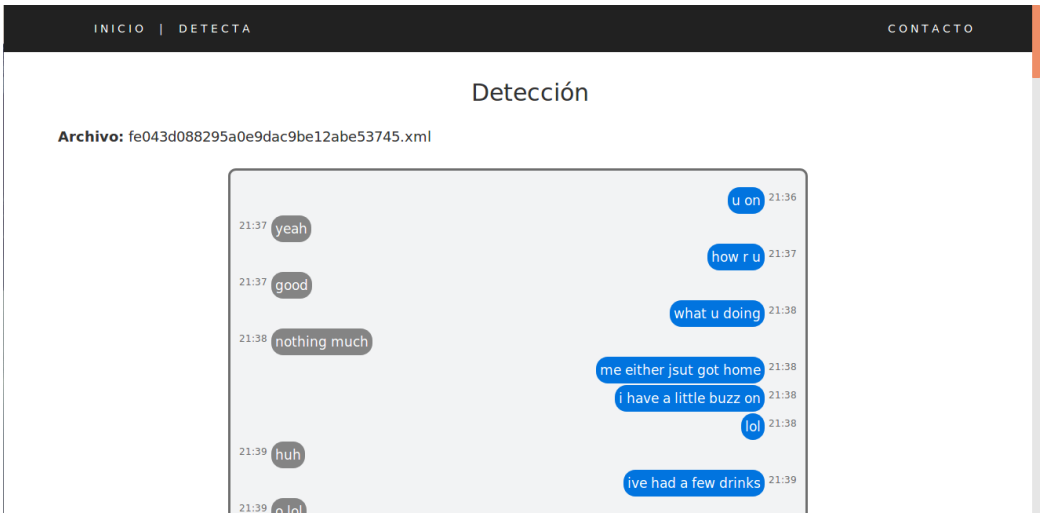

Figura 41: Página de visualización de la conversación del sistema web.

6. Conclusiones

En esta sección se hablará acerca de los objetivos que se lograron cumplir con el desarrollo de este trabajo, así como las conclusiones a las que se llegaron tras terminar este proyecto terminal.

Los objetivos planteados al principio de este proyecto se lograron completar satisfactoriamente, al haber desarrollado un modelo de predicción basado en clasificación supervisada de textos. Este modelo es capaz de identificar automáticamente líneas incriminatorias de acoso sexual dentro de una conversación.

Para poder generar el modelo de predicción, se realizó un análisis de las conversaciones de acoso sexual contenidas en el corpus utilizado. Este análisis permitió identificar aquellos atributos presentes en los textos, que proporcionan información útil para identificar líneas incriminatorias de acoso y que fueron de vital importancia para el modelo construido.

Asimismo, otro de los objetivos alcanzados fue implementar el modelo de predicción en un sistema web. Esto, permite visualizar tanto datos estadísticos de la conversación, como aquellas líneas presentes en la misma que son clasificadas como incriminatorias.

Estas visualizaciones proporcionadas por el sistema web, permiten al usuario conocer datos relevantes que se encuentran presentes en la conversación como las categorías de palabras que se utilizan en la conversación, la distribución de turnos, entre otros. De igual forma, permite identificar fácilmente y de manera óptima aquellas líneas de la conversación que pueden ser consideradas incriminatorias de acoso sexual.

Con esto se esperaba apoyar en la solución de casos de acoso sexual, ya que como se menciona en la justificación, solo muy pocos de estos quedan resueltos, debido a la compleja tarea de identificar evidencia incriminatoria dentro de una conversación de forma manual.

Sin embargo, es importante resaltar que este sistema no pretende sustituir el trabajo de los agentes de policía que se encargan de la solución de estos casos, por el contrario, se espera cooperar con estos proporcionándoles herramientas automáticas que puedan servir de apoyo en su trabajo.

Es por todo lo anterior que la motivación principal en este proyecto terminal fue desarrollar una herramienta computacional que facilite a un experto la identificación de evidencia de acoso sexual en conversaciones de chat.

Como trabajo futuro se planea realizar en un análisis más detallado de los atributos textuales con la finalidad de mejorar el desempeño del modelo generado. Así como también incorporar más características representativas para la visualización de la conversación, que puedan servir como evidencia incriminatoria en casos de acoso sexual, como por ejemplo que parte de las líneas de conversación (palabras o atributos) son más representativos.

Referencias

[1] Statista. *Ranking de las principales redes sociales a nivel mundial según el número de usuarios activos en abril de 2019 (en millones).* URL: https://es.statista.com/estadisticas/600712/ranking-mundial-de-redes-sociales-por-numero-de-usuarios/. (accessed: 28-09-2019).

[2] Datareportal. *Digital 2019: México.* URL: https://datareportal.com/reports/digital-2019-mexico. (accessed: 28-09-2019).

[3] INEGI. *Módulo sobre Ciberacoso (MOCIBA) 2017.* URL: https://t.co/oyG5W3VuNO. (accessed: 15.09.2019).

[4] Regeneración. *México primer lugar a nivel mundial en abuso sexual a menores: OCDE.* URL: https://regeneracion.mx/mexicoprimer-lugar-a-nivel-mundial-en-abuso-sexual-a-menores-ocde/. (accessed: 27.05.2018).

[5] Instituto De Estudios Sobre Sexualidad Y Pareja (INESSPA). *Abuso sexual infantil.* URL: http://www.inesspa.com/es/content/abusosexual-infantil. (accessed: 29.05.2018).

[6] Tom M Mitchell. *Machine learning*. 1997.

[7] Sebastián Maldonado y Richard Weber. «Modelos de selección de atributos para Support Vector Machines». En: *Revista Ingeniería de Sistemas* 26 (2012), págs. 49-70.

[8] Leticia C. Cagnina. «Representación de documentos». En: 2018.

[9] Hetal Bhavsar y Amit Ganatra. «A comparative study of training algorithms for supervised machine learning». En: *International Journal of Soft Computing and Engineering (IJSCE)* 2.4 (2012), págs. 2231-2307.

[10] Irina Rish y col. «An empirical study of the naive Bayes classifier». En: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. 2001, págs. 41-46.

[11] Jiawei Han, Jian Pei y Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[12] Leo Breiman. «Random forests». En: *Machine learning* 45.1 (2001), págs. 5-32.

[13] Payam Refaeilzadeh, Lei Tang y Huan Liu. «Cross-validation». En: *Encyclopedia of database systems* (2009), págs. 532-538.

[14] Rodrigo Alfaro, J. Cárdenas y Gastón Olivares Fernández. «Clasificación automática de textos usando redes de palabras». En: *Revista Signos* 47 (dic. de 2014), págs. 346-364. DOI: 10.4067/S0718-09342014000300001.

[15] Carlos-Emiliano González-Gallardo y col. «Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales». En: *Linguamática* 8.1 (2016), págs. 21-29.

[16] Shlomo Argamon y col. «Automatically profiling the author of an anonymous text.» En: *Commun. ACM* 52.2 (2009), págs. 119-123.

[17] Giacomo Inches y Fabio Crestani. «Overview of the International Sexual Predator Identification Competition at PAN-2012.» En: *CLEF (Online working notes/labs/workshop)*. Vol. 30. 2012.

[18] Hugo Jair Escalante y col. «Sexual predator detection in chats with chained classifiers». En: *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2013, págs. 46-54.

[19] Marius Popescu y Cristian Grozea. «Kernel Methods and String Kernels for Authorship Analysis.» En: *CLEF (Online Working Notes/Labs/Workshop)*. Rome. 2012.

[20] April Kontostathis y col. «Identifying Predators Using ChatCoder 2.0.» En: *CLEF (Online Working Notes/Labs/Workshop)*. 2012.

[21] Claudia Peersman y col. «Conversation Level Constraints on Pedophile Detection in Chat Rooms.» En: *CLEF (Online Working Notes/Labs/Workshop)*. 2012.

[22] Colin Morris y Graeme Hirst. «Identifying Sexual Predators by SVM Classification with Lexical and Behavioral Features.» En: *CLEF (Online Working Notes/Labs/Workshop)*. Vol. 12. 2012, pág. 29.

[23] Esaú Villatoro-Tello y col. «A Two-step Approach for Effective Detection of Misbehaving Users in Chats.» En: *CLEF (Online Working Notes/Labs/Workshop)*. 2012.