ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

ANÁLISIS DE LA VIOLENCIA DE GÉNERO DESDE LA PERSPECTIVA DE LA CIBERSEGURIDAD

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO/A EN CIENCIAS DE LA COMPUTACIÓN

PABLO ANDRÉS FLORES CHÁVEZ
HENRY EDUARDO YASIG QUILLUPANGUI

pablo.flores01@epn.edu.ec
henry.yasig@epn.edu.ec

DIRECTOR: MSC. PATRICIO XAVIER ZAMBRANO RODRIGUEZ
patricio.zambrano@epn.edu.ec

Quito, septiembre 2022

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Pablo Andrés Flores Chávez, bajo mi supervisión.

Ing. Patricio Xavier Zambrano Rodríguez

DIRECTOR DE PROYECTO

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Henry Eduardo Yasig Quillupangui, bajo mi supervisión.

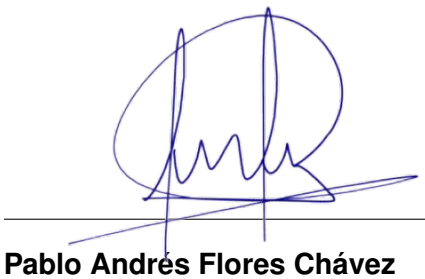
Ing. Patricio Xavier Zambrano Rodríguez

DIRECTOR DE PROYECTO

DECLARACIÓN

Yo, Pablo Andrés Flores Chávez, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

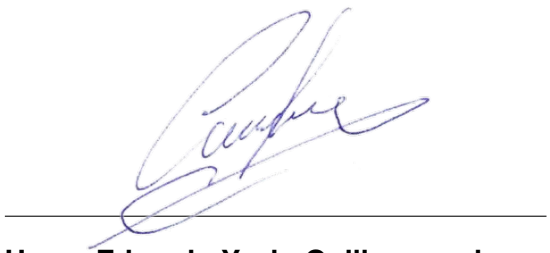


Pablo Andrés Flores Chávez

DECLARACIÓN

Yo, Henry Eduardo Yasig Quillupangui, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.



Henry Eduardo Yasig Quillupangui

DEDICATORIA

El presente trabajo de está dedicado a mi familia Raul Yasig, Virginia Quillupangui y mi hermano Fernando Yasig, por el apoyo incondicional que me han brindado a lo largo de mi vida universitaria, impulsándome a ser mejor persona cada día y a cumplir los sueños que me proponga en la vida.

A Camila Villagómez quien estuvo junto a mí en todo el proceso de mi vida universitaria, no fue sencillo culminar con éxito todo esté proceso, sin embargo, siempre fuiste positiva, confiable y amorosa, por esa razón en este momento se ha logrado culminar lo que un día me propuse.

A mis tíos, primos y amigos los cuales de igual manera han estado presentes en este proceso y me han apoyado incondicionalmente.

Henry Yasig

AGRADECIMIENTOS

La culminación de este objetivo tan importante en mi vida no hubiera sido posible sin el apoyo incondicional de la persona más importante de mi vida, mi madre, Patricia Chávez. Quien supo acompañarme en cada uno de los obstáculos que se presentaron en este camino y me dio la fuerza para sortearlos. También agradezco a mi abuelito, Alfonso Chávez, quien también fue parte fundamental de este proceso y sin el cual todo habría sido más complicado.

Pablo Flores

AGRADECIMIENTOS

Agradezco a Dios quien me ha guiado y me ha dado la fortaleza para seguir adelante y cumplir mis objetivos dentro de la universidad.
A la Escuela Politécnica Nacional por permitirme cursar la carrera en la facultad de ingeniería en sistemas, a sus docentes por brindarme sus conocimientos los cuales serán de gran ayuda en mi desempeño profesional.
Agradezco a mis padres, a mi hermano, primos y tíos por su amor, apoyo incondicional y ánimos para siempre salir adelante y cumplir con mi objetivo.

Henry Yasig

CONTENIDO

1 INTRODUCCIÓN

La violencia de género es un fenómeno que describe comportamientos violentos en la sociedad. El acoso online es ahora un concepto que gana cada día más atención, con presencia en varios canales de comunicación. Varios ataques como el bulliyng, la violencia de género, entre otros comparten el mismo patrón de hostigamiento.

El propósito de este trabajo es establecer relaciones conceptuales y procedimentales entre la violencia de género y el acoso en línea, dando como resultado el ciclo de vida de la violencia de género. Se procede a aplicar un proceso de minería de datos para el levantamiento de información sobre registros de experiencias de violencia de género. Dichas experiencias, permiten construir un patrón de ciclo de vida del ataque.

En los últimos años, los aspectos tecnológicos han evolucionado drásticamente [1], pero la violencia de género desde un punto tecnológico no ha crecido de la misma manera. Este hecho ha dificultado la detección y mitigación de esta. Una de las razones es la falta de confianza para hablar sobre el tema por parte de las víctimas, otra causa puede ser las políticas que tienen cada país para poder acceder a los datos relacionados.
Los autores en [2], consideran que estos aspectos pueden ser abordados para tener resultados más eficientes en investigaciones futuras. Los datos para la investigación pueden ser encontrados bajo diferentes fuentes de libre acceso, sin infringir las políticas de cada país y así minimizar el riesgo de los menores en relación con su privacidad e integridad. Dicho lo anterior se puede justificar la aplicabilidad del proceso de modelado propuesto a la violencia de género.

Las principales contribuciones de esta investigación se resumen a continuación:

- ❒ Este trabajo propone un ciclo de vida de la violencia de género desde la perspectiva de la seguridad informática.
- ❒ Se aplicará un modelo de IA no supervisado para agrupar las palabras de los textos en funcion de su contexto.

- ❒ Dotar de un contexto lingüístico e intención comunicacional a las distintas agrupaciones de palabras.
- ❒ Relacionar frases con diferentes etapas del ciclo de vida del ciberataque propuesto en círculos académicos.

1.1 ENTENDIMIENTO DEL PROBLEMA

1.1.1 Violencia de Género

Una de las definiciones más acertadas de la violencia de género fue propuesta por la ONU en 1995 [3] "Todo acto de violencia sexista que tiene como resultado posible o real un daño físico, sexual o psíquico, incluidas las amenazas, la coerción o la privación arbitraria de libertad, ya sea que ocurra en la vida pública o privada", de esta forma este ataque puede tomar diferentes formas como: físicas, psicológicas, sexual, social, entre otra.

Dicho esto, la violencia de género es la violencia que se desprende del hecho de ser hombre o ser mujer y es dirigida de un género a otro. Pero según varios estudios la violencia más frecuente es la del hombre hacia la mujer [4].

La violencia de género a ido creciendo dentro de las relaciones de pareja, formando de esta manera parte de la vida cotidiana de las mujeres y en ocasiones de hombres a lo largo del tiempo, esto se da porque las personas afectadas toman como un tabú hablar de la violencia que están sufriendo, es decir las mismas victimas lo consideran un asunto privado de pareja el cual no cuentan, ni hablan con nadie.

En la actualidad se ha ido avanzando en la sensibilización de esta problemática social, pero no se ha podido combatir del todo dado que aún hay actitud silenciosa por parte de las victimas [5], retrasando la investigación de este fenómeno social.

Uno de los factores por el cual se da la violencia de género es la normalización de las conductas por parte de las mujeres u hombres. Por parte de las mujeres, al tratarse de un problema cultural, muchas de ellas son socializadas en la aceptación de patrones de conducta abusivos sin darse cuenta de lo que hacen.

El ciclo de la violencia de género descrito por Walker [3] nos indica tres fases las cuales son: tensión, agresión y remisión. En la fase de remisión el atacante da a la víctima regalos o da signos de arrepentimiento, para mantener la amistad o confianza de la víctima. Lo que dificulta a la víctima detectar la situación y no saber cómo actuar luego de la agresión,

pensando que es una conducta propia de la pareja o amigo.

1.1.2 Estudios de la violencia de género

La violencia de género es un fenómeno nuevo en el campo de la investigación tecnológica. Luego de la realización del estado del arte se identificó una colección de trabajos académicos sobre el concepto de violencia de género en diversos campos de estudio de este fenómeno. Como resultado de la revisión, se obtuvo una agrupación sintetizada de las investigaciones en varios campos de estudio.

En la Tabla 1.1 se observa continuos aportes científicos en relación con el contexto social del ataque. Así mismo en menor medida, en las áreas de estudios técnicos, cultural y legal. Sin embargo, se evidencia que la cantidad de aportes con enfoque tecnológico aun es limitada.

Área de estudio	***Referencia***
Sociedad	[6], [7], [8], [9], [10], [11], [12]
Tecnología	[13], [14], [15], [16], [17], [9], [11], [18], [19]
Cultura	[8]
Leyes	[20]
Educación	[21], [22], [2], [23], [24]

Tabla 1.1: Clasificación por áreas de estudio de investigación sobre la violencia de género

1.1.3 Violencia de Género en la Sociedad

Este ataque es considerado uno de los problemas de alto impacto en la sociedad. Por esta razón han incrementado los índices de violencia de genero. Trayendo como consecuencia problemas psicológicos en personas de diferentes edades [6]. La violencia de género es una consecuencia de varios factores como: normas, valores, estereotipos y roles que se van aprendiendo en varios procesos de socialización de género.

El termino violencia de género no debe ser tratado como violencia contra las mujeres, hoy

en día son las principales afectadas, esto se da porque los hombres que son violentados no hablan por vergüenza, intimidación, entre otros factores. Dicho esto, existen algunos movimientos los cuales tratan de promover una sociedad que sea equitativa, dejando de lado los prejuicios basados en concepciones de género, es decir que no se corte la libertad propia de cada ser humano, de su independencia de sexo y orientación sexual [25].
Socialmente este ataque ha estado estigmatizado hacia las mujeres por varios factores como: su físico, fuerza entre otros. Trujillo y Pastor afirman que la violencia es un problema que amenaza el logro de la igualdad y por lo tanto es una violación de los derechos humanos y las libertades fundamentales [8]. Entonces se debe aclarar que los hombres no están exentos de dicha violencia.
La violencia de género puede darse de hombres a mujeres como de mujeres a hombres. Este problema va trascendiendo con el tiempo. Algunas de las causas por la que se da este problema son: la pobreza, falta de recursos económicos y falta de educación. Dejando que el problema transcurra de generación en generación por la falta de información.

1.1.4 Violencia de Género en la Tecnología

Hoy en día, l la tecnología ha hecho que muchos aspectos de nuestra vida diaria sean mucho más fáciles. Por todos los beneficios que trajo, también abrió la puerta a una multitud de nuevos inconvenientes [9]. Algunas personas utilizan el internet, los sistemas de mensajería y las redes sociales como una herramienta para demostrar comportamientos agresivos. Comportamientos que pueden ir desde el chantaje hasta el acoso, fomentando este ataque desde la coacción hasta el acoso.
Varios estudios sobre este problema han sido realizados abarcando el enfoque social. Sin embargo, Hinson afirma que desde el punto de vista de las tecnologías de información estos estudios aún son incipientes. Por esta razón muchos de los conceptos, las definiciones y la terminología del campo presenta inconsistencias [16].
Esta falta de estudios evita tanto la correcta diferenciación y clasificación de los ataques, como la medición y contabilización de estos. Con estos antecedentes, la tecnología de la información y la comunicación han propuesto varias técnicas que pueden medir y detectar ataques a través del aprendizaje automático y la minería de datos, tal como afirma Hidalgo-Leon [17].
Mediante estas técnicas se podrá obtener nuevos resultados que permitan entender la for-

ma en la que se realizan los ataques, e incluso predecirlos, tal como indica Rodríguez-Rodríguez [19]. Todo ello encaminado a mitigar los ataques de violencia de género. En el campo científico, varios autores toman como referencia la enfoque del ciclo de vida de Lockheed Martin [1] , quien desarrolló un modelo inicial de cadena de ciberataques. Bajo este criterio, el principal aporte de este estudio es el teórico/práctico definición del ciclo de vida de la violencia de género, desde el punto de vista de seguridad de la información.

1.1.5 Ciberacoso

El ciberacoso [26] se define como la intención sexual, como la conducta prejuiciada, realizada en Internet por adultos con el fin de ganarse la confianza de los menores y obtener gratificación sexual a través de imágenes eróticas o pornográficas proporcionadas por menores.

El internet ha sacado a la luz nuevos delitos que, por su gravedad, ameritan persecución penal.[26] Varios ataques como entrada ilegal a los sistemas informáticos, corrupción, ataque de denegación de acceso, nuevas tecnologías como: phishing, pharming, la difusión de pornografía infantil, ciberacoso, entre otras representan un verdadero desafío para las autoridades de varios países. Por esta razón algunas personas han encontrado formas nuevas y efectivas de llevar a cabo nuevos actos delictivos en la nueva tecnología[9]. Aprovechamos los vacíos legales que quedan al respecto debido a los constantes cambios en la tecnología y la falta de conocimiento sobre los peligros que implica el uso de la tecnología.

1.1.6 Definición de Objetivos

El objetivo de este estudio es establecer relaciones conceptuales y procedimentales de la violencia de género con criterios de la seguridad de la información y así definir un ciclo de vida de este fenómeno. Este objetivo general se fundamenta en los siguientes objetivos específicos:

- ❒ Investigar el contexto general de la violencia de género, actores y tendencias
- ❒ Analizar y establecer los patrones conductuales de los Atacantes.
- ❒ Establecer el comportamiento de los ataques asociados a la violencia de género (ciclo de vida).

1.1.7 Preguntas de Investigación

La finalidad del proyecto es crear un ciclo de vida de la violencia de género ligado a la seguridad de la información, para la realización de detecciones tempranas de ciber acoso en medios digitales. Por lo tanto, con el fin de lograr los objetivos antes mencionados, se plantearon las siguientes preguntas de investigación.:

RQ1: ¿Cómo se puede evaluar la naturaleza de la violencia de género en base a las experiencias recolectadas?

RQ2: ¿Cómo soporta el modelado de tópicos en la determinación de una solución viable al problema establecido?

RQ3: ¿Qué cantidad de información de las experiencias recolectadas resulta relevante para el estudio?

2 METODOLOGÍA

Este estudio, de carácter cualitativo, tiene como finalidad analizar a profundidad la violencia de género como una técnica de ataque relacionada a la seguridad de la información. Dado que ésta responde a un fenómeno social que tiene sus características intrínsecas propias y únicas [27]. Uno de los mecanismos de análisis de este fenómeno relacionado con el uso de la tecnología, fue propuesto por Leccese [27], misma motivación para estudiar el comportamiento humano a partir del análisis textual.
Existen varios mecanismos que permiten recabar información textual desde el internet [2] [19], en nuestro caso particular se recabaron experiencias o relatos directos de las víctimas, como casos-tipo. Cabe destacar que la información recolectada no está en relación directa con ellas. Puesto que la información recolectada contiene extractos de experiencias descritas de manera textual por las personas que han sido afectadas. Estos serán estandarizados mediante técnicas computacionales en un formato adecuado para su posterior análisis.

2.1 RECOLECCIÓN Y ANÁLISIS DE LOS DATOS

Toda la información que se obtuvo fue escrita, publicada y compartida por las victimas en diversos foros y páginas web. Para luego descargarlas mediante una herramienta de extracción de datos de código abierto Scrapy que tiene como principal objetivo extraer la información de determinados sitios web. Con la recolección de las muestras se obtuvo un paquete con 1616 historias de personas víctimas de violencia de género. Se registró cada historia con un número el cual nos sirve para tener un identificador de cada historia y, en los casos pertinentes la fecha de publicación de la historia en el sitio web. Toda esta información se descargó en un archivo .csv, el cual nos permite un manejo fácil de los datos extraídos. Con estos datos se procedió a la utilización de librerías para el manejo de textos, es decir, separar las stopwords y guardar las historias en otro archivo .csv. Luego se procede a lematizar los datos, es decir, las palabras que tenemos las vamos guardando en variables

para saber que categoría gramatical son como pueden ser: adjetivos, verbos y sustantivos para luego guardar estos datos en un archivo sin stopwords.

Creamos un diccionario con los datos obtenidos con el cual creamos un corpus de todas las palabras del texto, para contar la frecuencia con la que se va repitiendo cada termino dentro de un texto. Esta información se la ordena alfabéticamente y nos muestra el número de palabras que obtuvimos.

Con la información obtenida creamos una tabla por cada tópico. En nuestro caso se dividió en 9 tópicos, el cual nos indica el porcentaje de importancia que tiene una palabra dentro de cada uno. Para saber que tópico es el más dominante debemos saber el número de palabras que tiene cada historia para ello se lo analiza, categoriza y se le da contexto mediante la librería empath a cada uno. Con los resultados obtenidos procedemos realizar los gráficos de distribución de frecuencia. Para ello escogemos historias que tengan 100 a 200 palabras. No menos ni más porque afectan a los resultados y provocan ruido. Además, realizamos una nube de palabras con los resultados de cada tópico, para observar las palabras que son más relevantes.

2.1.1 Web Scraping

Web scraping se define como un mecanismo automatizado para la recolección de datos a partir de la web. Este define un cliente que realiza una solicitud a un servidor para la obtención de su contenido. El contenido será manipulado a través de parsers con el propósito de extraer información necesaria. Esté mecanismo se compone de una variedad de métodos y tecnologías para recopilar y extraer datos. Incluye el análisis de datos, parseo de lenguaje natural y seguridad de la información.[21]

2.1.2 CRIPS-DM

El proceso CRISP-DM se desarrolló por un consorcio compuesto inicialmente por Daimler Chryrler, SPSS y NCR. Su nombre CRISP-DM son las siglas de CrossIndustry Standard Process for Data Mining [24]. Según Wirth et al [23] el modelo CRISP-DM de minería de datos nos da una descripción del ciclo de vida del proyecto. Donde describe sus etapas, tareas y resultados. Adicionalmente, Huber et al. [22] lo caracteriza como un marco para la traducción de problemas comerciales en tareas de minería de datos. Además, esta meto-

dología permite la implementación de proyectos de minería de datos independientemente del campo de aplicación y la tecnología utilizada. La metodología CRISP-DM cuenta con 6 fases en Fig. 2.1

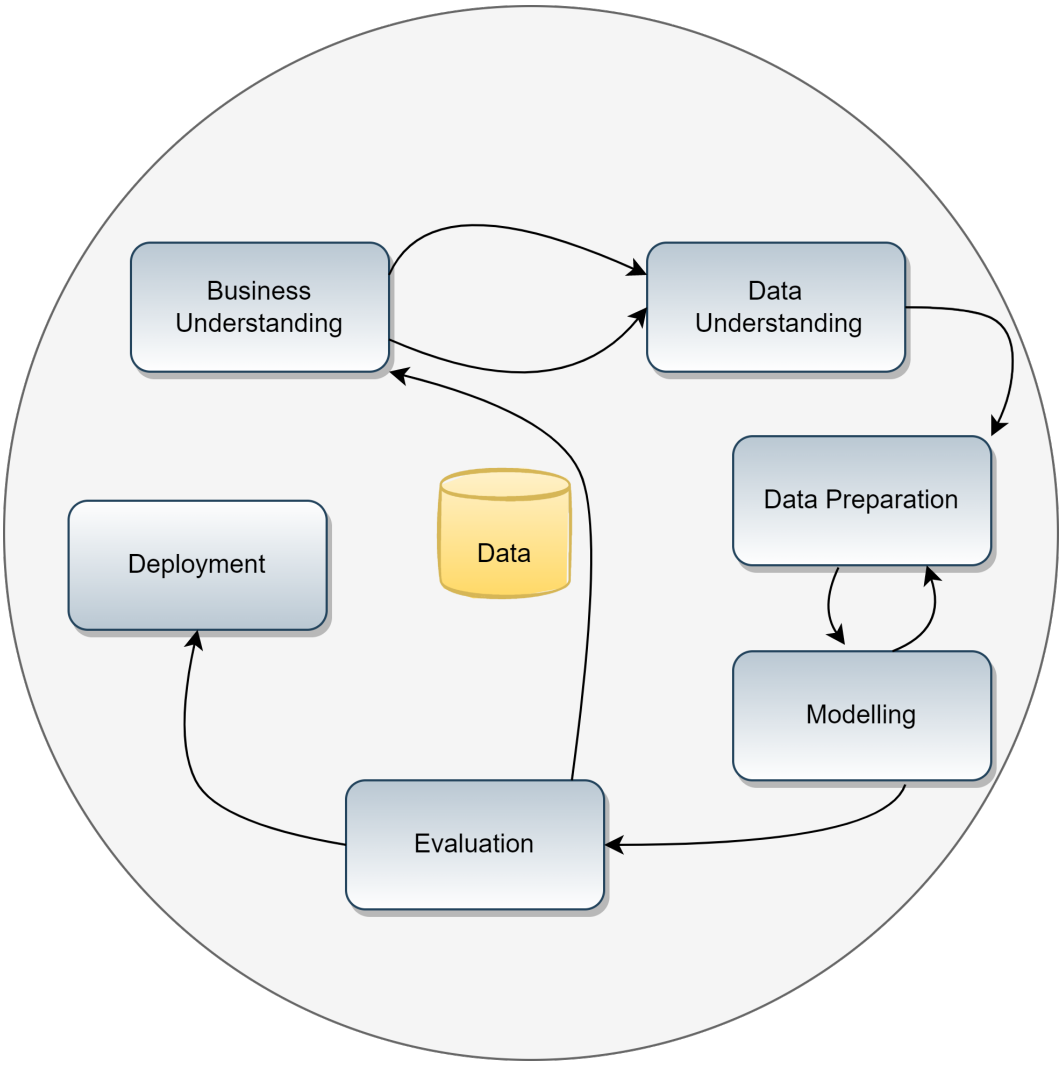


Figura 2.1: Fases del modelo CRISP-DM actual para la minería de datos

- La primera fase, trata sobre el entendimiento de los objetivos y requisitos del proyecto desde un punto de vista comercial. Se convertirá este conocimiento en un problema de minería de datos, preparando una etapa del proyecto preliminar diseñado para lograr los objetivos.

- La siguiente fase sobre la comprensión de datos empieza con la recolección inicial de datos y sus actividades de familiarización. Esta permite identificar problemas de calidad, descubrimiento de las primeras ideas, o la detección de patrones interesantes para la formulación de hipótesis.

- ❒ La fase de preparación de datos nos ayuda a cubrir las actividades con las cuales podremos construir los datos finales, es decir se realiza el tratamiento para el cambio de formato de la data de entrada en la fase de modelamiento, generando una estructura estandarizada.
- ❒ En el modelado se selecciona y aplican varias técnicas de modelado y calibra los parámetros a valores funcionales.
- ❒ En la etapa de evaluación los modelos obtenidos serán evaluados a fondo. Luego se revisan los pasos dados para construir un modelo que logre adecuadamente las metas propuestas.
- ❒ Finalmente, después de una exitosa evaluación viene la etapa de despliegue, donde se valida el impacto y relación con los objetivos planteados [22].

2.1.3 LDA

LDA es un algoritmo de inteligencia artificial no supervisado. La cual contiene una mezcla aleatoria de tópicos latentes. Que pueden ser extraídas como nociones abstractas que conforman los tópicos. Recibe como entrada un corpus conformado por una lista de documentos. Cada documento se compone de un vector de palabras, que identifican la presencia o ausencia en función del vocabulario. Y el vocabulario que corresponde a la lista de palabras existentes en todo el corpus. El resultado del modelo es una representación de los tópicos extraídos del corpus ingresado, este se valida con representaciones numéricas del peso de los tópicos sobre los documentos analizados.

2.2 DISEÑO DEL PROCESO DE INVESTIGACIÓN

La violencia de género es un problema social. Que conlleva muchos efectos negativos en las víctimas, y desde la perspectiva tecnológica este tema no ha sido muy abordado en estudios o investigaciones que ayuden a prevenir o detectar este ataque. Una de las metodologías aplicadas es: CRISP-DM, además de la utilización de varias cadenas de búsqueda que ayuden a encontrar información de historias, vivencias o narraciones de las víctimas o personas vinculadas al ataque. Tratando así de responder las preguntas de investigación antes planteadas.

Los datos encontrados fueron descargados y transformados mediante la técnica computacional conocida como scrapy, para luego proceder con la depuración de estos, mediante el uso de librerías de manejo de textos para poder separar las stopwords y lematizar las palabras. Mediante estos resultados, se tratará de determinar un ciclo de vida desde una perspectiva de la seguridad informática, relacionado al ataque de violencia de género. Con esta investigación se da a conocer más información sobre este ataque social, el cual servirá para futuras investigaciones del tema.

La secuencia narrativa se muestra en Fig. 2.2

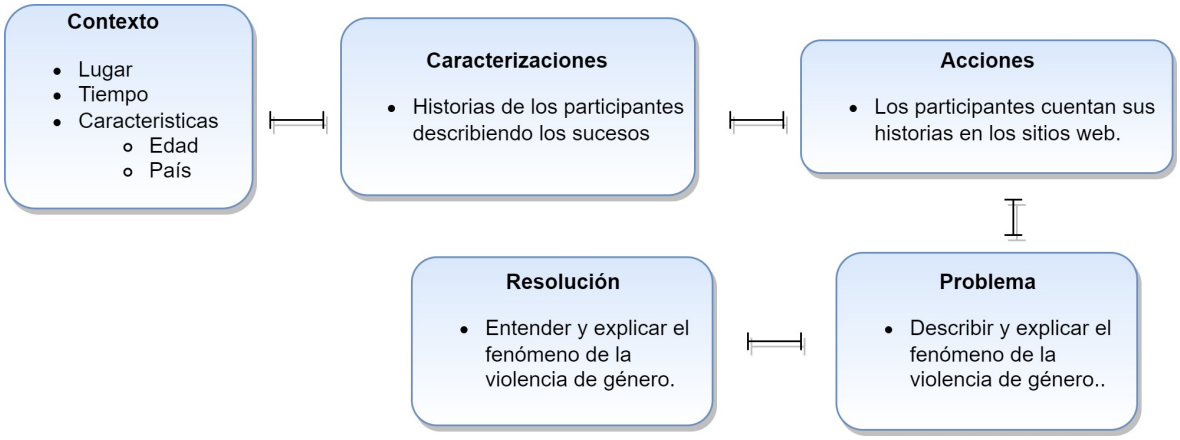


Figura 2.2: Fases de la secuencia narrativa

2.3 DESARROLLO DE LA METODOLOGÍA

2.3.1 Entendimiento del dominio del problema

En esta fase se revisa la literatura del dominio de la violencia de género, como podemos ver en la tabla 2.1. Así, se familiariza con los conceptos esenciales para tener una comprensión del tema y sus características. Se investiga la literatura existente referente a la violencia de género en diversas áreas de estudio.

Revisión de la literatura relacionada con el uso de tópicos	***Referencia***
Modelado de temas que utilizan contexto semántico para mejorar la clasificación de documentos	[28], [29], [30], [31], [32]
Modelado de temas que reconoce automáticamente patrones de expresión repetidos	[33]
Clasificación de texto usando LDA o variantes	[34], [35], [36], [37], [38],
Modelado de temas relacionado con Modelado de datos	[21], [22], [2], [23], [24]
Modelo de temas y experiencias de personas	[39], [40], [41], [42], [43]

Tabla 2.1: Modelado de temas - Revisión de la literatura

2.3.2 Entendimiento de los datos

Mediante la revisión de la literatura se observa que existen un gran grupo de personas que han sufrido experiencias donde han sido violentadas de distintas formas. Estas personas han sido vulneradas tanto verbal como físicamente, pasando por golpes o incluso violaciones.[16] Además, se pudo corroborar que el estudio de este problema desde una perspectiva tecnológica es prácticamente nulo.

Por esto, se ha visto la necesidad de proponer una investigación que aborde la temática desde el enfoque tecnológico, que sea soportado por herramientas que ayuden en el análisis de datos. Este enfoque permitirá elaborar un estudio sobre los patrones conductuales de las personas que ejercen violencia de género sobre otras. Con los resultados que se obtengan se podrá generar un recurso que permita mitigar e incluso prevenir estos ataques a futuro.

Para la elaboración de esta investigación se necesitan datos textuales que cuenten las experiencias de individuos que han sido violentados de esta forma. Estas experiencias se las obtiene mediante una búsqueda en internet y de varios sitios web de reputación conocida.

Para la búsqueda se utilizó Google. En este motor de búsqueda se buscó términos tanto en inglés como en español. Este fue un proceso iterativo que fue mejorando conforme se revisaba las entradas devueltas por el buscador. Al final se obtuvo cadenas de búsqueda que permitieron encontrar las publicaciones mencionadas en la tabla. 2.1.

Inicialmente se buscó algo sencillo como: "historias de violencia de género", "gbv stories.º "gv stories". Siendo "gbv"la abreviación usada en el idioma inglés para "gender based violencez "gv"la abreviatura en el mismo idioma para "gender violence". Los resultados de estas búsquedas fueron, casi en su totalidad, noticias que, para el propósito de la investigación, resultan irrelevantes.

Para mejorar los resultados se cambió la palabra "stories"por .experiences", utilizando las cadenas: .experiencias de violencia de género", "gbv experiences.º "gv experiences". Estas cadenas de búsqueda ya mostraban algunos resultados con experiencias reales, sin embargo, la cantidad de historias era reducida. Además, aun cuando los términos eran "gv.º "gbv", la mayor parte de los resultados únicamente mostraban resultados relacionados con "violence against women", también mencionado como "vaw", por sus siglas en inglés.

No se utilizaron abreviaturas para mejorar los resultados y se buscó específicamente: .experiencias de violencia de género", "gender based violence experiencesz "gender violence experiences". Con lo cual se pudo recuperar algunos sitios con contenido que más que contar historias, mostraba las definiciones y pequeños debates sobre el tema. Sin embargo, uno de los sitios encontrados fue de alta relevancia.

El sitio en cuestión es "Gender Links", el cual está web dedicado a recopilar historias de personas, en su generalidad mujeres, que han sufrido o a su vez han sido víctimas de este ataque y han querido contar su experiencia. El sitio permite compartir las historias para que más personas tomen fuerza y se animen a contar sus historias de abuso, además de encontrar apoyo en otras personas que han sufrido de la misma manera.

2.3.3 Preparación de los datos

En esta etapa se realiza el tratamiento de los datos para estandarizar el formato de estos, para su posterior uso (Aplicación de modelamiento de tópicos).

En Fig. 2.3, se describe el procedimiento para la obtención y preparación de datos.

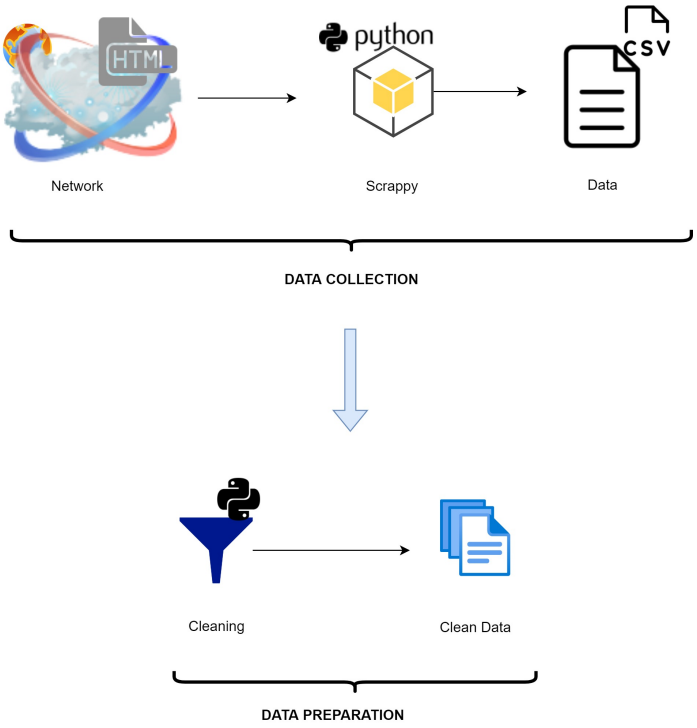


Figura 2.3: Preparación de datos

En la data collection tenemos el internet que es de donde encontramos las vivencias de cada persona, para luego utilizar Scrappy para la descarga de los mismo, dándonos como resultado una data en formato .csv el cual nos sirve para la preparación de los datos. En la preparación de los datos realizamos una limpieza de estas para eliminar stopwords, es decir eliminamos las palabras que no ayuden con la investigación. De esta manera se obtiene los datos limpios y listos para el modelo.

2.3.3.1 Herramientas utilizadas

La siguiente tabla 2.3 enumera el software, los recursos y las bibliotecas que se utilizan para implementar el proceso de minería de datos.

Tabla 2.2: Herramientas utilizadas

Herramienta	Descripción	Versión
Hardware		
Procesador	Intel Core i5-4310U CPU @ 2.00GHz 2.60GHz	N/A
RAM	Memoria instalada (RAM) 8GB	N/A
OS	OS 64bits, procesador x64	N/A
Software		
Windows	Sistema Operativo propietario de Microsoft.	Windows 10
Python	Lenguaje de programación, orientado al desarrollo de problemas de Inteligencia Artificial.	3.10.5 (64 bits)
Matlab	Lenguaje multiparadigma para computación numérica.	R2022a 9.12
Pandas	Librería para manipulación de datos	1.4.2
Chardet	Detector de codificación	4.0.0
Wordcloud	Librería de generación de gráficos (nubes de palabras).	1.8.2.2
Gensim	Librería para modelamiento de tópicos (Incluye el modelo LDA)	4.1.2
Lda2vec	Librería con modelo LDA modificado	0.16.10
Matplotlib	Librería para generación de gráficos genéricos.	3.5.1
Scrappy	Librería para extracción de datos de páginas web	2.6.1
Empath	Librería para categorización de términos.	N/A
spaCy y nltk	Librería para el procesamiento de lenguaje natural	N/A

Tabla 2.3: Herramientas utilizadas

2.3.3.2 Selección de datos

La violencia de género como ataque fue analizada y desarrollada desde la perspectiva de acoso, intimidación, problemas familiares, entre otros, ya que se evidenció que estos, son uno de los principales problemas del porque se da este ataque. Esta sección describe el análisis de los datos generados para el estudio, los detalles específicos se pueden visualizar en la tabla. 2.4

Violencia de Género	
Análisis previo de la literatura	si
Fuente de datos	Web
Tipo de datos	Charlas - Experiencias
Cantidad de datos	1616 Historias
Mecanismo de descarga	Descargas manuales, Script desarrollado en Python
Análisis de datos	Solo victimas (Experiencias)

Tabla 2.4: Selección de datos

2.3.3.3 Exploración de los datos

Mediante motores de búsqueda web y un proceso de refinamiento de consultas se indago repositorios que permitan encontrar experiencias de la violencia de género. Luego de una serie de filtrar resultados y refinamiento de las cadenas de búsqueda se encontró varios repositorios que pasaron a ser la base de conocimiento.

2.3.3.4 Obtención de datos

Mediante el Framework Scrappy, se realizó la descarga de los datos. Se definió el código de descarga, la lógica de depuración y extracción de los elementos, desde su formato base en HTML. Los datos se los almacenaron en un archivo base de formato .CSV. Al final se

obtuvo 1616 registros de diferentes experiencias de violencia de género. Fueron 13 sitios web de donde fueron recolectadas las experiencias las cuales se detallan en la siguiente tabla. 2.5

Sitio Web	Enlace
Women Against Abuse	www.womenagainstabuse.org/stories/
Nia Ending Violence	niaendingviolence.org.uk/get-informed/womens-stories/
Women Aid	www.womensaid.ie/help/stories.html
Survivor Stories Now	survivorstoriesnow.org/
RAINN	www.rainn.org/stories
Safe Steps	www.safesteps.org.au/survivor-stories/
Safe Lives	safelives.org.uk/new-views/real-life-stories
Domestic Shleters	www.domesticshelters.org/articles/true-survivor-stories
Healthy Place	www.healthyplace.com/abuse/rape
Take Back The Night	takebackthenight.org/category/child-abuse-survivor-stories
The Survivors Trust	www.thesurvivorstrust.org/blogs/survivor-stories?Take=20
Sayfty	sayfty.com/category/speak-our-stories/
Living Well	livingwell.org.au/from-men/stories-of-mens-experience/

Tabla 2.5: Fuentes de Obtención de Datos

2.3.3.5 Depuración de datos

El archivo .csv descargado con las experiencias de violencia de género, se lo manipula para poder realizar una limpieza de datos y generar un formato estandarizado, el cual será la data de entrada en el modelo de IA. Se realizaron las siguientes acciones:

- ❒ Creación de n-gramas.

- ❒ Eliminación de signos de puntuación y símbolos especiales.
- ❒ Eliminación de stop-words.
- ❒ Conversión de palabras en lexemas.

Utilizamos el modelo LDA que fue descrito anteriormente para trabajar con representaciones de n-gramas, por lo que se genera una representación de los documentos de hasta n = 3. Abstrayendo las diferentes combinaciones de términos, que contienen relevancia y significancia en el contexto. Es decir, combinaciones de palabras que juntas tienen un significado real, no serán reducidas a unigramas u omitidas en un significado diferente al usado en el contexto del documento. Posteriormente, las representaciones de los términos se convierten en un formato de lexema. Este permite reducir la representación de los términos a través de conjugaciones.
De esta forma los términos se concentran en enfatizar la definición y significado que estos representan. El formato final de los documentos ingresados es una lista de términos lematizados. Estos términos corresponden a un formato de "bag-of-n-grams"que representa a cada documento, y un vocabulario para todos los documentos. El formato de diccionario que fue generado se representa como asignaciones numéricas con una variable de frecuencia.

Dicha representación numérica, corresponde a la frecuencia de un determinado lexema en el documento correspondiente.

2.3.4 Modelamiento

En esta etapa se implementa el modelo de inteligencia artificial, que tiene como entrada los datos preparados para su uso. Y su resultado es el procesamiento del modelo.

2.3.4.1 Latent Dirichlet Allocatiion (LDA)

Se utilizó el modelo de Inteligencia Artificial no supervisado Latent Dirichlet Allocatiion (LDA). LDA [44] es muy utilizada en investigación y aplicaciones de programación neurolingüística. Este método utiliza distancias/ semánticas o similitudes entre términos para formar cadenas de palabras para encontrar palabras estrechamente relacionadas.

2.3.4.2 Preparación del modelo

La definición del modelo requiere el ingreso de las historias recolectadas para la distribución de tópicos, además de atributos de configuración del modelo:

- ❒ Topics: cantidad de tópicos a generar.
- ❒ Corpus: corresponde a la lista de documentos, en su representación de frecuencia de términos.
- ❒ Vocabulario: Lista de todos los términos existentes en el corpus.
- ❒ Número de tópicos: cantidad de tópicos a ser extraídos.
- ❒ Número de pases: Número de veces que el modelo trabajara sobre el corpus.
- ❒ Random state: valor de semilla para reproducibilidad del algoritmo.

2.3.4.3 Colección de documentos por tópicos

El modelo entrenado entrega, una representación de los documentos y sus términos en función de los tópicos extraídos. A cada elemento del vocabulario, se le ha asignado el tópico de mayor relevancia al cual corresponde. Así como un valor de peso sobre el tópico asignado.

A partir de cada termino asignado a los tópicos sugeridos, se procede a separar dichos términos en subgrupos por cada tópico. Los grupos corresponden a una representación de términos relacionados. El subgrupo pasara a asignarse una categoría lexicográfica mediante la librería “Empath”. Esta librería recibe una cadena de texto, la cual se conforma a partir de los términos de cada tópico que LDA género.
Posteriormente, los términos de entrada generan nuevas categorías léxicas. Definiendo una frecuencia de categorías por tópico, a partir de los lexemas de los subgrupos y la frecuencia. Corresponde a las nuevas categorías obtenidas a partir de los términos del tópico. Las categorías no corresponden al mismo texto de lexemas, del tópico que LDA clasifico. Sino a una relación de los lexemas iniciales, en los tópicos que “Empath” ofrece.

2.3.4.4 Porcentaje de contribución de los tópicos en los documentos

A partir de los documentos, y la asignación de peso de sus términos a través de los diferentes tópicos generados, se clasifica los documentos en los tópicos. Para ello, a partir de la lista de términos se obtiene el de mayor peso o relevancia. Este representara el tópico dominante del documento.
A partir de la clasificación de documentos por tópicos, se observa la naturaleza del documento o el contenido de este. A partir de su texto se puede inferir el tipo de documento que constituye el tópico asignado. Por lo tanto, los documentos de menor relevancia tendrán un esquema o naturaleza similar al de mayor relevancia.

2.3.5 Evaluación

En esta etapa se evalúan los resultados obtenidos del modelo seleccionado, aplicando diferentes combinaciones de la data preparada. Dado que de esta manera se podrá encontrar un resultado de mayor calidad y cohesión en sus tópicos.

2.3.5.1 Definición de solución de compromiso

Para una correcta implementación del modelo fue necesario realizar pruebas variando la data de entrada de éste. Para estas pruebas fue necesario dividir en categorías las 1616 historias que se obtuvieron. La categorización se hizo en función de la cantidad de palabras de cada una de las historias después de ser procesadas. La cantidad de palabras seleccionada para el rango de categorización fue de 500 palabras. Con esta configuración se obtuvieron 11 categorías. Tal como se observa en la siguiente tabla 2.6.

Número de Palabras	Cantidad de historias
0 - 500	36
500 - 1000	62
1000 - 1500	163
1500 - 2000	379
2000 - 2500	442
2500 - 3000	239
3000 - 3500	144
3500 - 4000	73
4000 - 4500	38
4500 - 5000	12
5000+	28
TOTAL	**1616**

Tabla 2.6: Número de palabras y cantidad de historias recolectadas

Además, podemos ver la proporción que hay con las historias recolectadas por el número de palabras en Fig. 2.4.

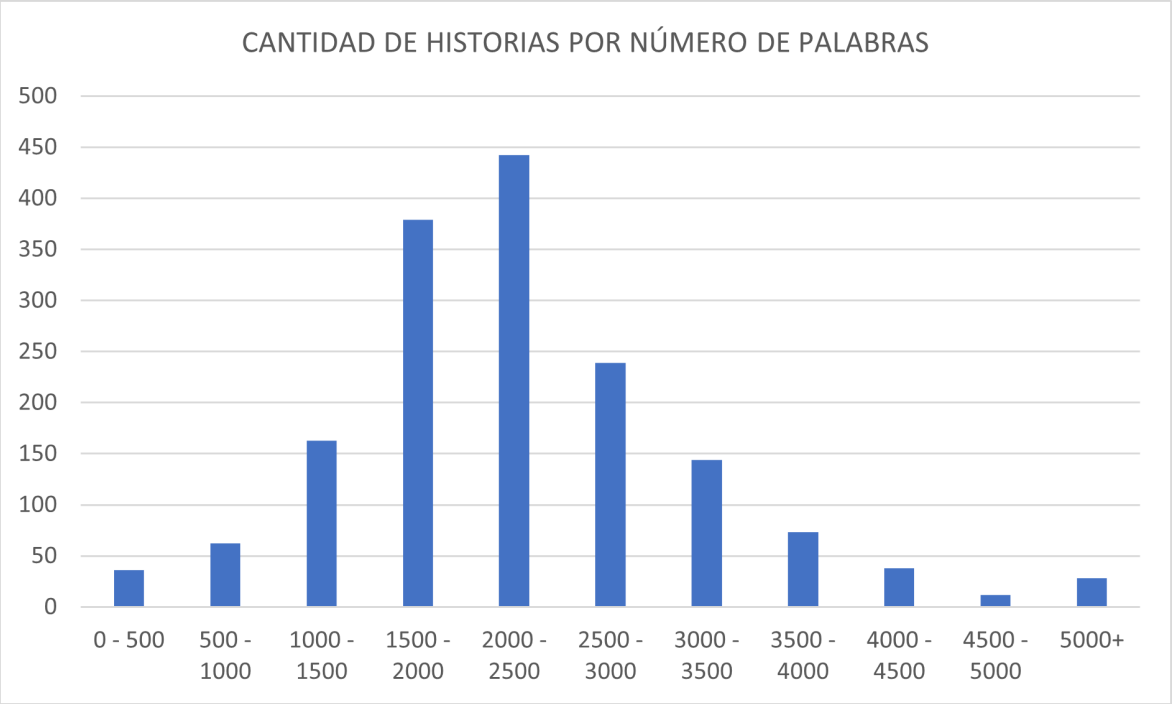


Figura 2.4: Cantidad de Historias - Número de palabras

2.3.5.2 Comparación de resultados

Con los datos ya categorizados se procedió a probar el modelo con cada una de las categorías para que el modelo sea el más eficiente y que sirva para cumplir con los objetivos de este trabajo investigativo. Las pruebas consistieron en medir el desempeño del modelo con los datos determinados. Para esto se generó el gráfico de la perplejidad contra el coste computacional requerido para analizar las historias. Con las gráficas fue posible determinar el mejor conjunto de datos de entrada para el cual el modelo se comporta de manera equilibrada.

Las pruebas iniciaron con historias de texto largo (gran cantidad de palabras). Para la primera prueba se consideraron las 28 historias más largas, es decir, las correspondientes a la categoría de más de 5000 palabras. El gráfico generado con estos datos mostraba un incremento si cesar de la perplejidad. Por esta razón, se desechó esta configuración. Tal como se observa en la Fig. 2.5.

Para la siguiente prueba se aumentó la cantidad de historias, pero se disminuyó la cantidad de palabras por cada una de estas. Se corrió el programa con un total de 73 historias. Estas correspondían a la categoría de entre 3500 y 4000 palabras. El comportamiento del gráfico

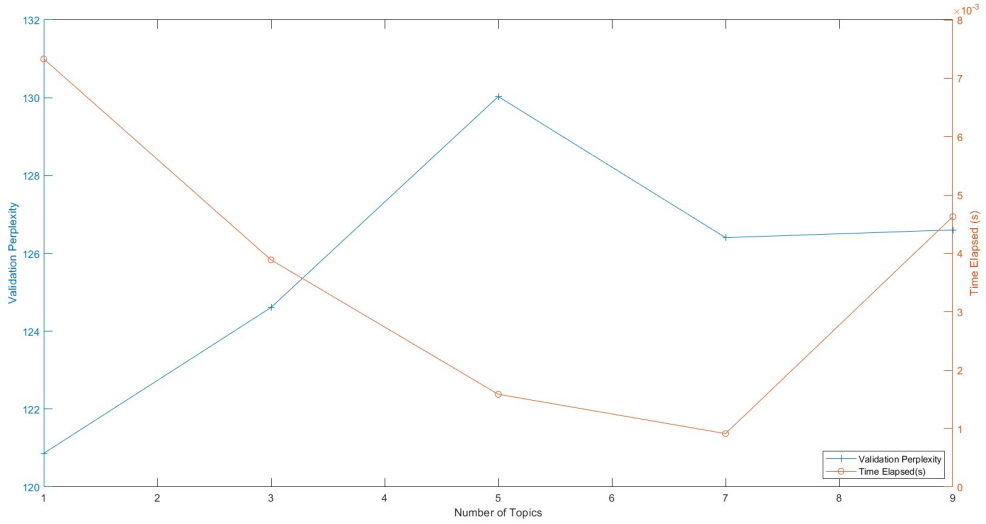


Figura 2.5: 28 historias - 5000+ palabras

fue similar al anterior. La perplejidad tenía una tendencia creciente hasta el infinito, tal como se observa en la siguiente figura Fig. 2.6 . Por esta razón por la que se desechó esta data.

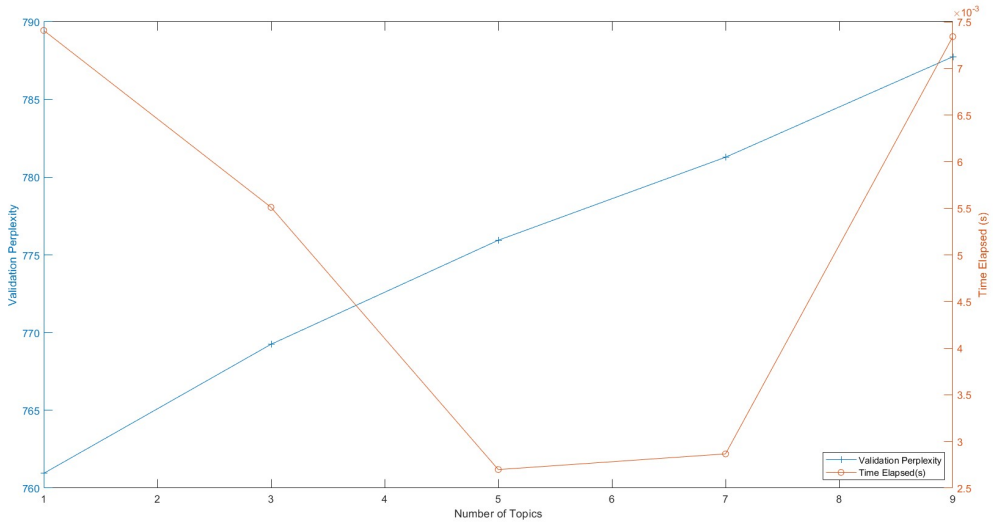


Figura 2.6: 73 historias - 3500-4000 palabras

Después se buscó analizar una gran cantidad de historias que tuvieran menos palabras. Fueron seleccionadas 442 historias, correspondientes a todas aquellas que tuvieran entre 2000 y 2500 palabras. Sin embargo, el gráfico Fig. 2.7, de estas seguía mostrando información no relevante para los objetivos planteados en la investigación.

Posteriormente se buscó reducir tanto la cantidad de historias como la cantidad de palabras. Así fue como se corrió el programa con un total de 36 historias. Estas fueron co-

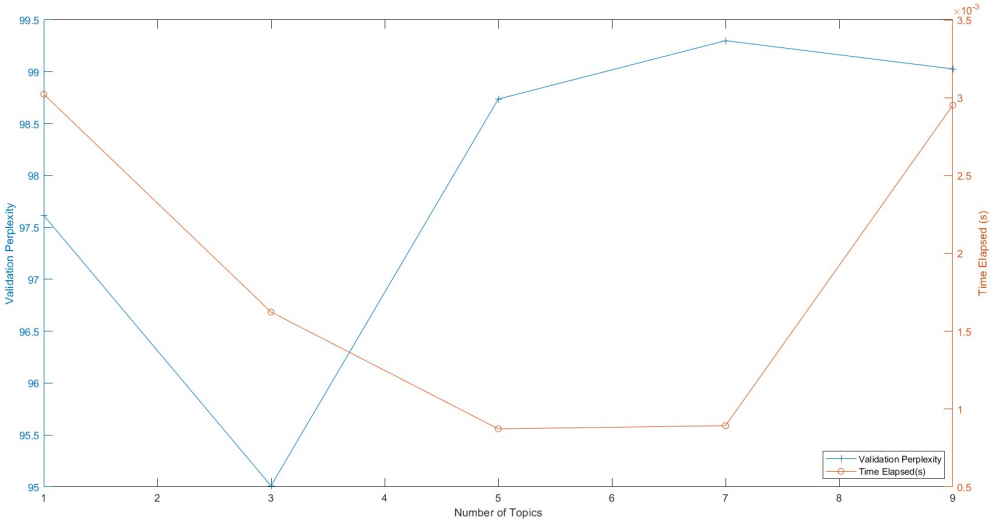


Figura 2.7: 442 historias - (2000-2500) palabras

rrespondientes a todas las historias con menos de 500 palabras. De igual manera que en el caso anterior, el gráfico mostraba una relación inversa a la requerida. Mientras el costo computacional disminuía, la perplejidad aumentaba. Tal como se observa en la Fig. 2.8

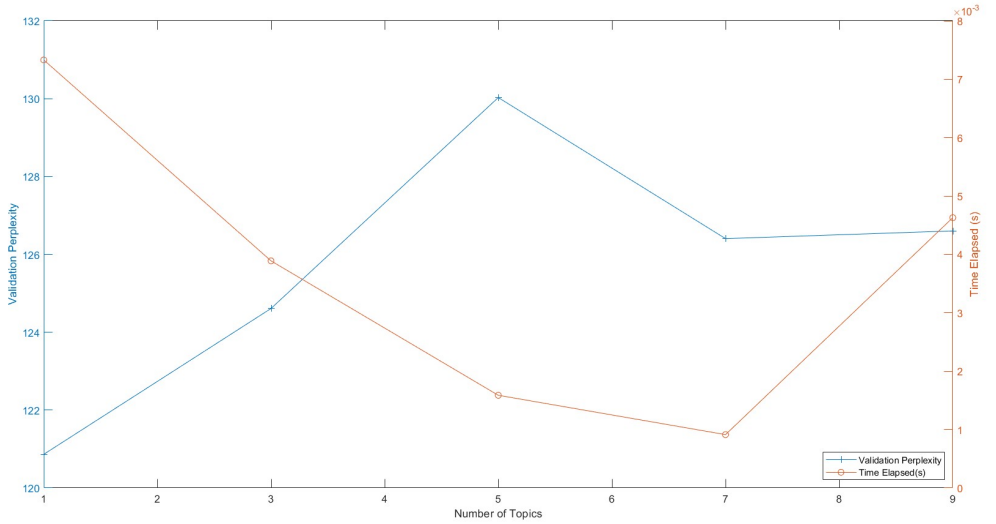


Figura 2.8: 36 historias - (0-500) palabras

Posteriormente se utilizó las 15 historias con menos palabras. Estás historias tenían entre 130 y 385 palabras. El gráfico generado para estas historias no presentaba un aporte significativo para la investigación, por esta razón fue desechado. De manera similar ocurrió con el gráfico generado utilizando as 20 historias más cortas (entre 130 y 468 palabras). Ambos gráficos mostraban una relación en la cual tanto la perplejidad como el costo computacional tenían una tendencia decreciente., tal como se observa en las figuras

Fig. 2.9 y Fig. 2.10.

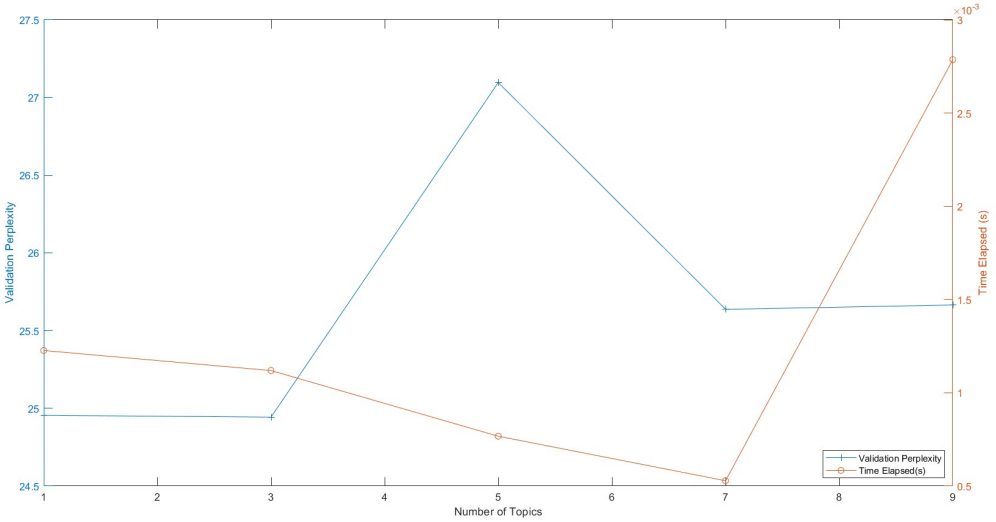


Figura 2.9: 15 historias más cortas

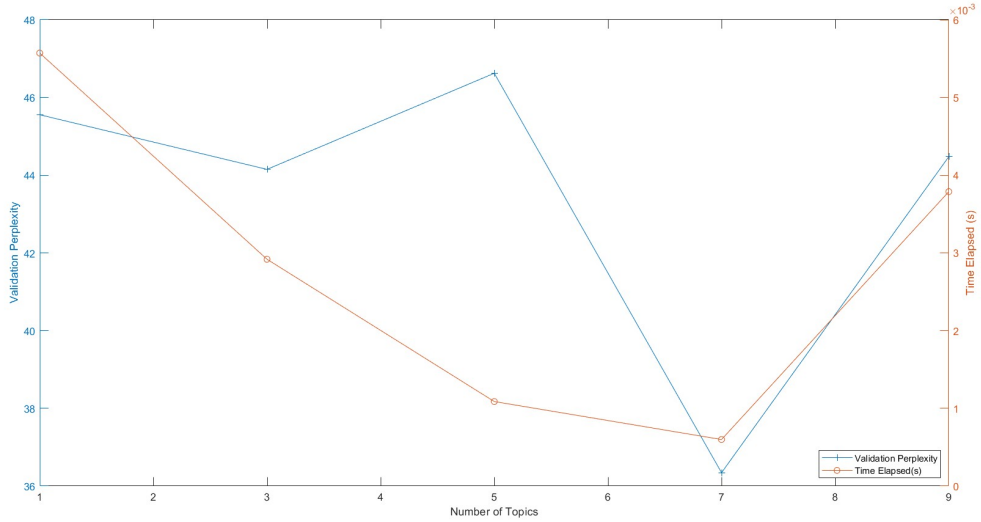


Figura 2.10: 20 historias más cortas

Finalmente, se probó aumentando la cantidad de historias hasta 50. El gráfico obtenido con las 50 historias de menor tamaño ya mostraba información que podía ser utilizada para mejorar el desempeño del modelo. Tal como se observa en Fig. 2.13, las curvas del costo computacional y la perplejidad mostraban una relación coherente entre sí. Mientras que la perplejidad iba disminuyendo, el costo computacional iba aumentando.

Adicionalmente, estas curvas se intersecan en un valor muy cercano a 5, correspondiente al número de tópicos ideal a ser relacionado con las fases de un ataque de violencia de

género.

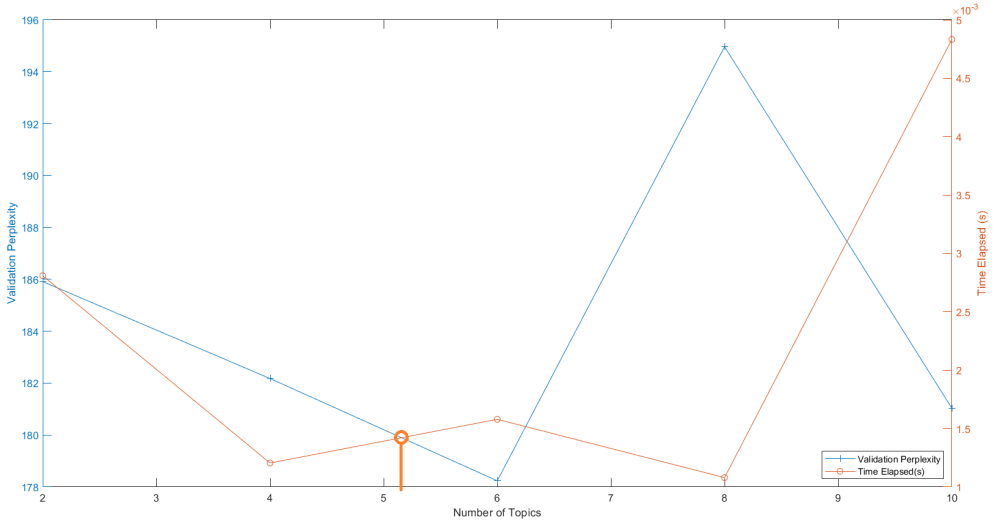


Figura 2.11: 50 historias más cortas

Comparando las gráficas obtenidas con los distintos conjuntos de datos de entrada se escogió el escenario de las 50 historias más cortas para realizar el análisis. Con estos datos se generaron 5 tópicos, cada uno con su nube de palabras como se observa en Fig. 2.12. El modelo LDA requirió un conjunto de datos para realizar la evaluación de perplejidad vs costo computacional como se pudo apreciar en la comparación de modelos, de esta forma determinamos un número óptimo de temas.

Figura 2.12: Nube de palabras por tópicos

Detallando en la tabla 2.7 las palabras de cada tópico encontrado.

Tópico	Palabras en la Nube
1	Leave, domestic, violence, abuse, victim
2	Ago, need, year, rape, friend
3	People, keep, tell, happen, late
4	Three, woman, myself, help, every
5	Know, take, life, nigh, time

Tabla 2.7: Tópicos de la nube de palabras

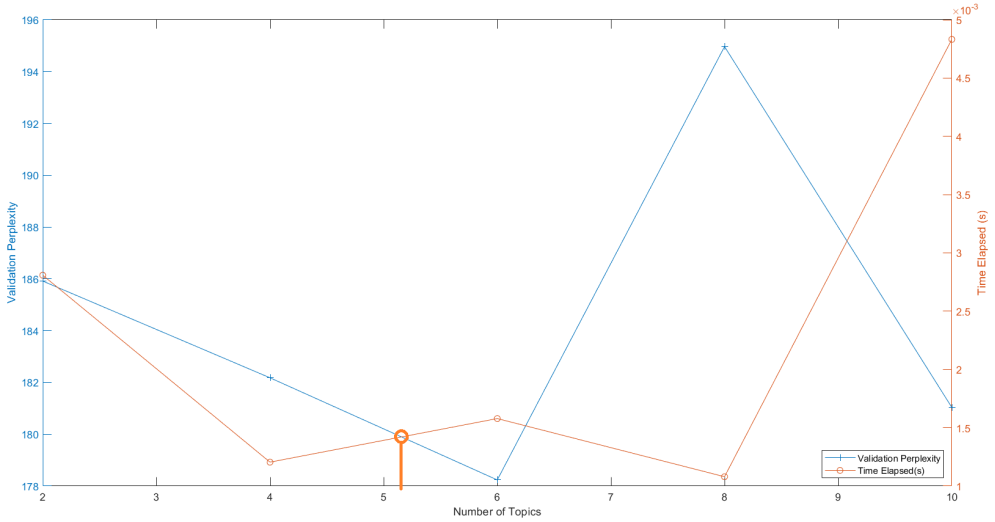


Figura 2.13: Perplejidad de la nube seleccionada

2.3.5.3 Tópicos creados

Para determinar el número de temas de esta investigación, se procedió a elegir un rango de valores que tiene varios números de temas, y en el transcurso del análisis determinar una solución óptima basada en la perplejidad y el tiempo de procesamiento en la aplicación del modelo. En el caso que el número de temas sea alto, se puede elegir un valor más bajo

para agilizar el proceso de ajuste y así determinar un numero apropiado de temas [45]. El rango considerado para determinar que una solución sea optima comienza con dos temas propuestos por Lancaster y puede llegar a ocho temas propuesto por Mandiant, BSI y Sdapt [46] El resultado del valor optimo determino que el número de tópicos adecuados para el análisis del ciclo de vida de la violencia de género es seis como se ilustra en Fig. **??**. Como puede verse, la perplejidad y el tiempo transcurrido para este número de temas es razonable, pero se optó por escoger 5 tópicos.

Los temas obtenidos mediante el modelo LDA son los descritos en Fig. 2.14

Figura 2.14: Ciclo de vida de la violencia de género

3 RESULTADOS Y DISCUSIÓN

3.1 CONTRASTE DE TÓPICOS CON MODELOS DE CICLO DE VIDA DE ATAQUES

A través del proceso descrito en este documento, hemos podido obtener las historias de varias personas que han sido víctimas de violencia de género en algún momento de sus vidas. Estos datos se obtuvieron mediante un proceso automatizado en Python y luego se analizaron mediante el modelado de temas.

Las experiencias adquiridas nos permitieron profundizar en el tema de la violencia de género. Además, mediante el modelado de temas, pudimos generar cinco tópicos diferentes y agrupar palabras según el contexto y la intención comunicativa. Además de distinguir nuevas etapas, estos grupos nos han permitido establecer vínculos con diferentes etapas del ciclo de vida del ciberataque propuesto en el ámbito científico.

Luego de realizar el modelo de investigación y obtener los tópicos del ciclo de vida del ataque se analizaron los nombres de cada tópico:

- Tópico 1: Ejecución del ataque [1], donde el atacante realiza el ataque a su víctima,
- Tópico 5: Persistencia del ataque, donde el atacante busca la manera de llegar a la víctima.
- Tópico 4: Explotación de la relación [47], donde el atacante ya realiza el ataque y tiene contacto con la víctima.
- Tópico 2: Acción de la víctima violentada, la víctima busca ayuda.
- Tópico 3: Afectación de la víctima, donde la víctima tiene problemas a casusa de la violencia de género.

3.2 DISCUSIÓN

3.2.1 ¿Cómo se puede evaluar la naturaleza de la violencia de género en base a las experiencias recolectadas?

La naturaleza de este fenómeno se la evaluó mediante tres aspectos principales los cuales son: los datos recolectados, el modelo y los resultados. Dentro de los datos se optó por analizar textos escritos por personas que han sufrido violencia de género. Los textos recolectados son historias de personas que han sido víctimas de este fenómeno. Dándonos las experiencias vividas sin diferenciar una edad especifica. Los datos fueron descargados desde distintos sitios web que fueron descritos en la sección 2.3.3.4.

Estos documentos fueron los que sirvieron como insumo en la evaluación del modelo de aprendizaje automático. Una vez analizadas las historias se las pasó por un modelo LDA para revisar y obtener patrones distintivos en cada una de ellas. Posteriormente cada una de estas historias fue preprocesada. El preprocesamiento de los datos consistió en poner todas las palabras en minúsculas, eliminar los signos de puntación, Stopwords o palabras vacías y de esta forma obtener únicamente las partes más relevantes de cada una de ellas. Al tener una cantidad relativamente grande de historias, se realizó un análisis para encontrar la cantidad de estas que aportaban información relevante para la investigación. Para esto se estudió las gráficas de la perplejidad contra el costo computacional arrojadas por el programa. Se empezó probando con las historias que más cantidad de palabras tenían, sin embargo, no se obtuvieron resultados que aportaran a cumplir el objetivo de este trabajo. Luego se probó con una cantidad de palabras más baja y se obtuvieron mejores resultados utilizando únicamente las 50 historias más cortas.

Fue así como se consiguieron cinco tópicos en funcion de su carácter lingüístico. Junto con cada uno de ellos se generó además una nube de palabras. Con estas nubes de palabras fue posible establecer las fases correspondientes dentro del ciclo de vida del ataque. La evaluación de los resultados se la realizó mediante el agrupamiento de los textos en categorías. Además, se analizaron semejanzas y diferencias con las fases descritas en la sección, 2.3.5.3 para entender y evaluar, mediante cada tópico, las características principales de cada una de las fases obtenidas.

3.2.2 ¿Cómo soporta el modelado de tópicos en la determinación de una solución viable al problema establecido?

El modelado de tópicos es una técnica que nos permitió definir criterios que permitan describir un conjunto de experiencias para el análisis. El modelo determina los tópicos que se encuentra implícitamente incluidos en los documentos y los agrupa con el propósito de comprender y resumir las historias en palabras de gran importancia. Varios estudios como [2] [45] lo soportan mediante trabajos previos, los cuales estudiaron los datos que se tiene en análisis de textos. En nuestro caso los resultados fueron de mayor cantidad y de esta manera se afirma las comunicaciones previas permitiendo crear un ciclo de vida. En esta investigación se optó por usar nubes de palabras ya que estas hicieron más sencilla la tarea de asociar cada uno de los tópicos a una fase específica del ciclo de vida de un ciberataque. De esta manera, con los tópicos y las de nubes de palabras obtenidas a partir del modelamiento es posible describir el ciclo de vida de la violencia de genero. Por esta razón, el modelado de tópicos se presenta como una herramienta de mucha ayuda para tratar de comprender un fenómeno como la violencia de género. Con los resultados obtenidos en la sección 2.3.5.3, se pueden presentar estrategias que permitan disminuir los casos de violencia de género en la tecnología mediante la identificación de las fases encontradas.

3.2.3 ¿Qué cantidad de información de las experiencias recolectadas resulta relevante para el estudio?

En esta investigación se reunieron 1616 experiencias de personas que, en algún momento de su vida, han sido víctimas de la violencia de género. Algunas de las víctimas han sufrido ataques por mucho más tiempo que otras, por lo cual algunas historias son más largas. Para tener un conocimiento más amplio de la información que tenían las historias se contó el número de palabras de cada una. Con esto se obtuvo que la historia más corta tenía únicamente 131 palabras, mientras que las historias más largas llegaban a tener más de 5000.

Al tener una gran cantidad de historias, se procedió a agruparlas en función del número de palabras que componen cada historia. Se conformaron un total de 11 categorías, en intervalos de 500 palabras por historia. Con este agrupamiento se empezó a realizar pruebas

del modelo para encontrar cuales historias daban un resultado más eficiente que otro.

Así, se probó con distintas combinaciones tanto de historias, como del tamaño de estas. En primera instancia se pensó que mientras más palabras tuvieran las historias más información relevante tendrían estas. Sin embargo, al momento de realizar la ejecución del programa se fue evidenciando lo contrario. Mientras más palabras tenían las historias, los resultados eran más deficientes. La cantidad tan grande de palabras en lugar de aportar algo relevante a la investigación, entorpecía la ejecución del programa y a su vez los resultados que este ofrecía.

Así fue como se procedió a realizar pruebas con las historias que menos palabras tenían. Las pruebas del modelo cuando se utilizaron historias con un número significativamente menor de palabras fueron más reveladoras. Por tal razón se decidió trabajar con las 50 historias de menor tamaño. Con estas historias, los tópicos y las nubes de palabras generados ya presentaban resultados importantes y con sentido. Con estos resultados fue posible mapear los tópicos a las fases del ciclo de vida de ciber ataques.

4 CONCLUSIONES

Como parte de nuestra investigación, se propuso un modelo de ciclo de vida del fenómeno de la violencia de género desde una perspectiva de seguridad de la información. Esto se realizó a través de un proceso de minería de datos, sirviéndonos de la técnica CRISP-DM. Esta técnica plantea las fases y actividades para realizar minería de datos.

Los datos fueron obtenidos mediante una búsqueda exhaustiva en varios sitios web. Para ello se realizó un proceso de refinamiento de cadenas de búsqueda, con la finalidad de encontrar repositorios o bases de datos que contengan experiencias relevantes de personas que han sufrido violencia de género. Una vez que se obtuvieron sitios web de confianza se procedió a descargar las experiencias de estos mediante un programa en Python. Este programa se encargó de recorrer la página, descargarla y presentarla en un archivo de valores separados por comas (.CSV). Si bien los sitios web consultados corresponden a instituciones u organizaciones reconocidas que trabajan o tienen programas para combatir la violencia de género, al ser historias de personas anónimas, no podemos afirmar a ciencia cierta que estas sean auténticas. Puesto que no se mostraban nombres reales, identificaciones o correos electrónicos de las víctimas con los cuales corroborar la información. Es por esta razón, que dentro de nuestra investigación no se demostró la autenticidad de las 1616 experiencias que se obtuvieron. No obstante, estas historias sirvieron para obtener resultados deseados correspondientes al ciclo de vida de un ataque.

Con esta cantidad de historias se procedió a agruparlas en funcion del número de palabras de estas. Por ello es importante mencionar que la cantidad de información que se tenga en la base de datos si llega a ser relevante. Puesto que las historias que contienen una cantidad de palabras demasiado grande generan ruido entorpeciendo el análisis. Por esta razón las historias muy extensas no son de utilidad en el método planteado.

Otro de los criterios importantes dentro de nuestra investigación fue el idioma en el cual se encontraban las experiencias. El idioma seleccionado fue el inglés, dado que hay mucha más información y es un lenguaje más universal. Por lo que se descartaron historias encontradas en español, francés y portugués.

Para este estudio cada una de las experiencias fue depurada. Esta depuración consistió en la eliminación de signos de puntuación, Stopwords y una correcta tokenización de las palabras. El análisis realizado después del preprocesamiento de las historias permitió generar nubes de palabras correspondientes a cinco tópicos que describen el ciclo de vida de la violencia de género.

El ciclo de vida obtenido se compuso de cinco fases. En la primera fase, correspondiente al tópico 1, se realiza la ejecución del ataque. Esta fase es donde el atacante realiza de manera efectiva cualquier tipo de violencia contra la víctima. En la segunda fase, correspondiente al tópico 5, se observa una persistencia en el ataque. Es decir, que el atacante ejerce de manera reiterada cualquier tipo de violencia contra la víctima. En algunos casos los ataques pueden ir subiendo en intensidad. En la tercera fase, correspondiente al tópico 4, se observa una explotación de la relación. En esta fase el atacante tiene un mayor contacto con la víctima. En la cuarta fase, correspondiente al tópico 2, se evidencian las acciones tomadas por la victima violentada. Es aquí donde se visualiza que la víctima busca ayuda. Por último, la quinta fase, correspondiente al tópico 3, muestra las afectaciones que tuvo la víctima. En esta fase la victima tiene problemas, que pueden ser físicos o psicológicos, a causa de la violencia sufrida.

De este modo se logró establecer el comportamiento general de los ataques asociados a la violencia de género. Teniendo en cuenta cómo y cuándo empieza un ataque, en la gran mayoría de los casos, de forma persistente y repetitiva, hasta las consecuencias que estos ataques tienen en la vida de las víctimas.

Una de las principales limitantes dentro de nuestra investigación es la falta de información sobre experiencias describiendo al atacante, dado que el modelamiento fue realizado únicamente desde la perspectiva de la víctima. Puesto que la información desde el punto de vista del atacante es escasa.

Finalmente, uno de los principales desafíos en el camino hacía nuevos estudios de casos es obtener datos relevantes para el área de investigación. Por ello, se recomienda difundir este tipo de investigaciones dentro de la comunidad científica.

5 REFERENCIAS BIBLIOGRÁFICAS

[1] E. M. Hutchins, M. J. Cloppert, R. M. Amin et al., «Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains,» *Leading Issues in Information Warfare & Security Research*, vol. 1, n.º 1, pág. 80, 2011.

[2] P. Zambrano, J. Torres, Á. Yánez, A. Macas y L. Tello-Oquendo, «Understanding cyberbullying as an information security attack—life cycle modeling,» *Annals of Telecommunications*, vol. 76, n.º 3, págs. 235-253, 2021.

[3] F. Expósito, M. Moya et al., «Violencia de género,» *Mente y cerebro*, vol. 48, n.º 1, págs. 20-25, 2011.

[4] M. Banchs, «Violencia de género,» *Revista venezolana de análisis de coyuntura*, vol. 2, n.º 2, págs. 11-23, 1996.

[5] A. J. Y. Garcıa, «La violencia contra las mujeres: conceptos y causas,» *Barataria. Revista Castellano-Manchega de Ciencias Sociales*, n.º 18, págs. 147-159, 2014.

[6] I. Valor-Segura, F. Expósito y M. Moya, «Victim blaming and exoneration of the perpetrator in domestic violence: The role of beliefs in a just world and ambivalent sexism,» *The Spanish journal of psychology*, vol. 14, n.º 1, págs. 195-206, 2011.

[7] B. Zurbano Berenguer, I. Liberia Vayá y B. Campos Mansilla, «Concepto Y Representación De La Violencia De Género: Reflexiones Sobre El Impacto En La Población Joven (Concept and Representation of Gender-Based Violence: Reflections About the Impact on Young People),» *Oñati socio-legal series*, vol. 5, n.º 2, 2015.

[8] M. Trujillo Cristoffanini e I. Pastor-Gosálbez, «Violencia de género en estudiantes universitarias: Un reto para la educación superior,» *Psicoperspectivas*, vol. 20, n.º 1, págs. 83-94, 2021.

[9] M. Á. Verdejo Espinosa et al., *Ciberacoso y violencia de género en redes sociales: análisis y herramientas de prevención*, 2015.

[10] M. ElSherief, E. Belding y D. Nguyen, «# notokay: Understanding gender-based violence in social media,» en *Eleventh international AAAI conference on web and social media*, 2017.

[11] J. Xue, J. Chen y R. Gelles, «Using data mining techniques to examine domestic violence topics on Twitter,» *Violence and gender*, vol. 6, n.º 2, págs. 105-114, 2019.

[12] F. R. Pastene, S. Niklander, J. Irarrázaval e I. Luengo, «Gender violence: media treatment and analysis of the Nabila Rifo case in La Cuarta and Las Últimas Noticias,» en *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2020, págs. 1-6.

[13] H. Khatri e I. Abdellatif, «A Multi-Modal Approach for Gender-Based Violence Detection,» en *2020 IEEE Cloud Summit*, IEEE, 2020, págs. 144-149.

[14] Á. M. Martınez, M. d. M. S. Márquez, A. B. B. Martın, M. d. M. M. Jurado, M. del Carmen Pérez-Fuentes y J. J. G. Linares, «Revisión del uso de las nuevas tecnologıas para la intervención en violencia de género en parejas de adolescentes,» *European Journal of Child Development, Education and Psychopathology*, vol. 1, n.º 1, págs. 63-73, 2013.

[15] D. Á. Garcıa, J. C. N. Pérez, L. Á. Pérez, A. D. González, C. R. Pérez y P. G. Castro, «Violencia a través de las tecnologıas de la información y la comunicación en estudiantes de secundaria,» *Anales de Psicologıa/Annals of Psychology*, vol. 27, n.º 1, págs. 221-231, 2011.

[16] L. Hinson, J. Mueller, L. O'Brien-Milne y N. Wandera, «Technology-facilitated gender-based violence: What is it, and how do we measure it?,» 2018.

[17] P. Hidalgo-León, «Gender Violence's Models and Discrimination-aware Data Mining.,» en *EKAW (Doctoral Consortium)*, 2018.

[18] O. M. Cumbicus-Pineda, T. E. Abad-Eras y L. A. Neyra-Romero, «Data Mining to Determine the Causes of Gender-Based Violence against Women in Ecuador,» en *2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)*, IEEE, 2021, págs. 1-6.

[19] I. Rodrıguez-Rodrıguez, J.-V. Rodrıguez, D.-J. Pardo-Quiles, P. Heras-González e I. Chatzigiannakis, «Modeling and forecasting gender-based violence through machine learning techniques,» *Applied Sciences*, vol. 10, n.º 22, pág. 8244, 2020.

[20] G. Coll-Planas, G. G.-R. Moreno, C. M. Rodrıguez y L. Navarro-Varas, «Cuestiones sin resolver en la Ley integral de medidas contra la violencia de género: las distinciones entre sexo y género, y entre violencia y agresión,» *Papers: revista de sociologıa*, págs. 187-204, 2008.

[21] R. Lawson, *Web scraping with Python*. Packt Publishing Ltd, 2015.

[22] S. Huber, H. Wiemer, D. Schneider y S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model,» *Procedia Cirp*, vol. 79, págs. 403-408, 2019.

[23] R. Wirth y J. Hipp, «CRISP-DM: Towards a standard process model for data mining,» en *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Manchester, vol. 1, 2000, págs. 29-40.

[24] A. Azevedo y M. F. Santos, «KDD, SEMMA and CRISP-DM: a parallel overview,» *IADS-DM*, 2008.

[25] K. A. Zayas, «Violencia de género: pandemia de la sociedad,» *Estudios del Desarrollo Social: Cuba y América Latina*, vol. 3, n.º 2, págs. 87-98, 2015.

[26] V. P. Galence, «El ciber-acoso con intención sexual y el child-grooming,» *Quadernos de criminologıa: revista de criminologıa y ciencias forenses*, n.º 15, págs. 22-33, 2011.

[27] R. Hernández-Sampieri, C. Fernández Collado, P. Baptista Lucio et al., *Metodologıa de la investigación*. McGraw-Hill Interamericana México, 2018, vol. 4.

[28] S. Jameel, W. Lam y L. Bing, «Supervised topic models with word order structure for document classification and retrieval learning,» *Information Retrieval Journal*, vol. 18, n.º 4, págs. 283-330, 2015.

[29] D. Peng, D. Guilan y Z. Yong, «Contextual-LDA: a context coherent latent topic model for mining large corpora,» en *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, IEEE, 2016, págs. 420-425.

[30] D. Liu, Y. Zeng, Y. Luo, H. Pang y X.-H. Wu, «Window-Based Topic Model for HDP,» en *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, IEEE, 2019, págs. 70-75.

[31] M. Allahyari y K. Kochut, «Automatic topic labeling using ontology-based topic models,» en *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2015, págs. 259-264.

[32] J. Bai, L. Li y D. Zeng, «Activating topic models from a cognitive perspective,» en *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, IEEE, 2016, págs. 55-60.

[33] A. Trabelsi y O. R. Zaıane, «A joint topic viewpoint model for contention analysis,» en *International Conference on Applications of Natural Language to Data Bases/Information Systems*, Springer, 2014, págs. 114-125.

[34] E. Laoh, I. Surjandari y L. R. Febirautami, «Indonesians'Song Lyrics Topic Modelling Using Latent Dirichlet Allocation,» en *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, IEEE, 2018, págs. 270-274.

[35] S. ElShal, M. Mathad, J. Simm, J. Davis e Y. Moreau, «Topic modeling of biomedical text,» en *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2016, págs. 712-716.

[36] Y. Luo y H. Shi, «Using lda2vec topic modeling to identify latent topics in aviation safety reports,» en *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, IEEE, 2019, págs. 518-523.

[37] S. Mifrah y B. L. El Habib, «Semantic Relationship Study between Citing and Cited Scientific Articles Using Topic Modeling,» en *Proceedings of the 4th International Conference on Big Data and Internet of Things*, 2019, págs. 1-8.

[38] C. Zhai, «Probabilistic topic models for text data retrieval and analysis,» en *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, págs. 1399-1401.

[39] J. Y. A. PÚBLICA, J. E. HERRIZAINGO y H. A. SAILA, «MUJERES VÍCTIMAS DE VIOLENCIA DE GÉNERO: VIVENCIAS Y DEMANDAS,»

[40] C. Gonzalez, «Testimonios de mujeres supervivientes de violencia de género,» 2017.

[41] A. 3. Noticias, «Los duros testimonios de dos víctimas de la violencia de género: Aprendí a conducir por la N-1 mientras él me golpeaba,» 2018.

[42] M. SEMITIEL, «Testimonios que matan,» 2018.

[43] M. Rojas, «El testimonio de Ana Bella, víctima de maltrato: Nunca te separarás de mí, lo nuestro es amor o muerte,» 2020.

[44] D. M. Blei, A. Y. Ng y M. I. Jordan, «Latent dirichlet allocation,» *Journal of machine Learning research*, vol. 3, n.º Jan, págs. 993-1022, 2003.

[45] P. Zambrano, J. Torres, L. Tello-Oquendo et al., «Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach,» *IEEE Access*, vol. 7, págs. 142 129-142 146, 2019.

[46] B. I. Messaoud, K. Guennoun, M. Wahbi y M. Sadik, «Advanced persistent threat: New analysis driven by life cycle phases and their challenges,» en *2016 International conference on advanced communication systems and information security (ACOSIS)*, IEEE, 2016, págs. 1-6.

[47] K. D. Mitnick y W. L. Simon, *The art of deception: Controlling the human element of security*. John Wiley & Sons, 2003.