DEPARTAMENTO EN CIENCIAS DE LA COMPUTACIÓN E INTELIGENCIA ARTIFICIAL

UNIVERSIDAD DE GRANADA

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

APLICACIÓN DE TÉCNICAS DE MINERÍA DE TEXTO AL ESTUDIO DE LA VIOLENCIA CONTRA LA MUJER

AUTORA

Stephanie Elizabeth Mora Andrade

DIRECTORES

María del Carmen Pegalajar Jiménez

María Amparo Vila Miranda

Editor: Universidad de Granada. Tesis Doctorales
Autor: Stephanie Elizabeth Mora Andrade
ISBN: 978-84-1195-334-4
URI: https://hdl.handle.net/10481/92541

DEDICATORIA

A mi familia, cuyo apoyo y aliento inquebrantable ha sido la luz que me ha guiado a lo largo de este viaje. Vuestra fe en mí ha alimentado mi determinación para alcanzar este hito, y estoy infinitamente agradecida por su amor y comprensión.

A mis directoras, cuya experiencia y orientación han dado forma a mi trayectoria académica y ha enriquecido mi comprensión en el campo de la minería de texto. Vuestra tutoría ha sido inestimable para dar forma a la dirección de esta tesis y a mis futuras actividades.

A las innumerables personas que han participado en mi investigación. Su disposición y apoyo ha sido fundamental para el desarrollo de este trabajo.

A mis amigos y colegas, que me han proporcionado compañerismo, ánimo y un sentimiento de comunidad durante los altibajos de este esfuerzo académico.

Y, por último, a la propia búsqueda del conocimiento. Esta tesis es un tributo a la búsqueda incesante de la comprensión, y la dedico a la exploración interminable que nos impulsa a ampliar los límites de la comprensión humana.

RESUMEN

La Violencia Contra la Mujer (VCM) es un problema de carácter social que está presente en muchos países, convirtiéndose en un fenómeno de alto alcance que requiere atención y un amplio estudio para así concienciar a la sociedad sobre el impacto y las consecuencias que genera. Así mismo, la creciente ola de casos de violencia hace evidente la necesidad urgente de reconocer su importancia y asegurar los principios relativos a la igualdad, seguridad, libertad, integridad y dignidad de todos los seres humanos. Lo antes expuesto motivó a que se realice la presente investigación doctoral para estudiar y analizar las formas y diferentes patrones que envuelven la VCM. De manera que, mediante la aplicación de diferentes técnicas de minería de texto y aprendizaje automático sobre una gran variedad de noticias recolectadas de diversos periódicos digitales obtuvimos información valiosa y relevante que nos ha proporcionado una visión profunda de este fenómeno social latente a nivel mundial.

En esta investigación se propone el uso de técnicas de Minería de Texto como: Clasificación de texto, Modelado de temas y Reglas de Asociación para realizar un estudio de la VCM tomando como fuente artículos de violencia extraídos de periódicos digitales. Primeramente, en esta investigación se empleó técnicas de Raspado Web para obtener la colección de documentos a ser estudiados. Una vez obtenida la colección de documentos se sugiere realizar lo siguiente: clasificación del texto en los diferentes tipos de violencia que sufren las mujeres, de la misma forma, mediante la aplicación de técnicas de modelado de temas, se generarán e identificarán temas latentes dentro de la colección de documentos. Finalmente, con la aplicación de minería de reglas de asociación se propone el estudio de los diferentes atributos y patrones que involucran la violencia contra la mujer.

Esta propuesta consiste en el desarrollo de los siguientes puntos:

- Inicialmente, para poder llevar a cabo esta investigación se comenzó con la recopilación de noticias públicas por periódicos digitales. Previo a la recopilación fue necesario realizar un estudio de las diferentes estructuras de páginas web, de

modo que, se pudiera identificar en qué nodo de la estructura HTML se encontraba el texto requerido, a fin de, poder definir sobre qué nodos de información haremos la petición, y así, poder obtener el texto específico de cada una de las noticias.

- Con las técnicas de raspado web se pudo recopilar 7000 noticias (documentos de texto) en formato no estructurado.

- Posteriormente, se procedió a realizar el procesamiento de la colección de documentos. Este proceso presentó algo de complejidad debido a que el objeto de estudio es texto que puede contener cientos de palabras, donde cada palabra representa un atributo, de modo que, los documentos a estudiar son de gran dimensionalidad. Este tipo de dato sin estructura es más complejo de estudiar dado que muchos de los atributos presentes en el texto no generarán valor a la investigación o incluso podrían afectar al buen funcionamiento de los algoritmos de aprendizaje automático. Para reducir el impacto de este problema se aplicó un proceso de procesamiento de texto que permitió la selección de las características más relevantes dentro de cada uno de los textos recopilados para el estudio.

- Para identificar y determinar los tipos de violencia con los que se clasificarían los documentos se realizó un estudio de casos e investigaciones sobre VCM que permitieran determinar los tipos de violencia que sufren las mujeres. De aquí se obtuvieron 3 tipos de violencia: Física, Sexual y Psicológica, las cuales pueden estar relacionadas y presentes en un solo hecho o documento, por lo que, se optó por una clasificación multiclase.

- Para la detección de temas latentes se utilizaron técnicas de modelado de temas, en este estudio se aplicó el algoritmo Asignación Latente de Dirichlet, conocido con sus términos en inglés como “Latent Dirichlet Allocation” (LDA). Como resultado, se obtuvo una lista de temas junto con sus 15 términos más representativos, así mismo, se pudo detectar ciertas características sobre la VCM. A continuación, se determinó las noticias más relevantes dentro de cada tema, y mediante las palabras más frecuentes se pudo construir etiquetas de identificación.

- Finalmente, en el proceso de minería de reglas de asociación se realizó un estudio de las diferentes características que pueden involucrar un acto de violencia. Entre estas tenemos: el tipo de víctima, el tipo de agresor, los motivos, el arma empleada, el tipo de violencia, si existen heridas en el cuerpo o si la víctima murió o no. A partir del procedimiento descrito anteriormente se aplicaron reglas de asociación sobre una colección de 7000 documentos. Posteriormente, fue necesario realizar una reducción de dimensionalidad, debido a que cada documento puede contener una gran cantidad de palabras. La razón de realizar esta reducción fue los recursos informáticos que consumen la aplicación de modelos de reglas de asociación en documentos de gran dimensionalidad. Así mismo, el uso de atributos poco importantes en la generación de reglas de asociación podría generar resultados dudosos en las dependencias de los atributos.

Los resultados obtenidos en el proceso de desarrollo de esta investigación fueron favorables demostrando que las técnicas de minería de texto son herramientas de gran utilidad en el estudio de la Violencia Contra la Mujer, estas técnicas nos permitieron estudiar hechos reales de violencia y obtener información que antes era desconocida. Finalmente, se pudo evidenciar la gravedad y el gran alcance que tiene la VCM, además, de observar la necesidad de aplicar medidas que ayuden a la erradicación de este fenómeno universal que acecha a miles de mujeres y niñas a nivel mundial.

ABSTRACT

Violence Against Women (VAW) is a social problem that is present in many countries and has become a far-reaching phenomenon that requires attention and extensive study in order to raise awareness of its impact and consequences. Furthermore, the growing wave of cases of violence makes evident the urgent need to recognize its importance and to ensure the principles of equality, security, freedom, integrity and dignity of all human beings. The mentioned before motivated the present doctoral research to study and analyze the forms and different patterns that surround VAW. By applying different text mining and machine learning techniques to a wide variety of news collected from different digital newspapers, we obtained valuable and relevant information that has provided us with a deep insight into this latent social phenomenon at a global level.

This research proposes the use of Text Mining techniques such as Text Classification, Topic Modelling and Association Rules to carry out a study of VAW taking as source articles of violence extracted from digital newspapers. Firstly, in this research, we used Web Scraping techniques to obtain the collection of documents to study. Once we obtained the collection of documents, the proposal is as follow; classification of the text into the different types of violence suffered by women, in the same way, through the application of topic modelling techniques, generate and identified latent topics within the collection of documents. Finally, with the application of association rule mining, the study of the different attributes and patterns involving violence against women is proposed.

This proposal consists of the development of the following points:

- Initially, in order to carry out this research, we began with the collection of news published by digital newspapers. It was necessary to carry out a study of the different web page structures in order to identify in which node of the HTML structure the required text is located. In order to be able to define on which information nodes we will make the request and thus be able to obtain the specific text of each of the news items.

- Using web-scraping techniques was possible to collect 7000 news items (text documents) in unstructured format.

- Subsequently, the document collection was processed. This process was complex because the text could contain hundreds of words, with each word representing an attribute. Thus, the documents to study were of high dimensionality. This type of unstructured data is more complex to study; many of the attributes present in the text will not generate value to the research or could even affect the proper functioning of the machine learning algorithms. To reduce the impact of this problem, we applied a text processing process that allowed the selection of the most relevant characteristics within each of the texts collected for the study.

- In order to identify and determine the types of violence to classify the documents, we studied research on VAW to determine the types of violence suffered by women. We detected three types of violence: Physical, Sexual and Psychological, which can be related and present in a single event or document, so we chose a multi-class classification.

- For the detection of latent topics, we used modelling techniques. In this study, we applied the Latent Dirichlet Allocation (LDA). As a result, we obtained a list of topics and their 15 most representative terms, as well as certain characteristics of VAW. Then, the most relevant news within each topic was determined, and by extracting the most frequent words, we constructed identification tags for the generated topics.

- Finally, in the association rule mining process we study the different characteristics that may be involved in an act of violence. These include; the type of victim, the type of aggressor, the motives, and the weapon used, the type of violence, whether the victim has wounds on the body or whether the victim died or not. Based on the procedure described above, we applied association rules on a collection of 7000 documents. Subsequently, it was necessary to perform a dimensionality reduction, because each document can contain a large number of words. The reason for this reduction was the computing resources consumed by the application of association rule models on high-dimensional documents. In

addition, the use of unimportant attributes in the generation of association rules could generate dubious results in the attribute dependencies.

The results obtained in the process of developing this research were favorable, demonstrating that text-mining techniques are very useful tools in the study of violence against women, as they allowed us to study real facts of violence and obtain information that was previously unknown. Finally, it was possible to demonstrate the seriousness and the great scope of VAW, as well as to observe the need to apply measures that help to eradicate the universal phenomenon that stalks thousands of women and girls worldwide.

Tabla de Contenido

INDICE DE FIGURAS

INDICE DE TABLAS

LISTA DE ACRÓNIMOS

[VCM] VIOLENCIA CONTRA LA MUJER

[MT] MINERÍA DE TEXTO

[IA] INTELIGENCIA ARTIFICIAL

[AA] APRENDIZAJE AUTOMÁTICO

[SVM] MÁQUINA DE SOPORTE DE VECTORES (SUPPORT VECTOR MACHINE)

[RF] ARBOLES ALEATORIOS (RANDOM FOREST)

[NB] REDES BAYESIANAS (NAÏVE BAYES)

[DT] ARBOLES DE DECISION (DECISION TREE)

[ANN] RED NEURONAL ARTIFICIAL (ARTIFICIAL NEURONAL NETWORK)

[PLN] PROCESAMIENTO DE LENGUAJE NATURAL

[RI] RECUPERACIÓN DE INFORMACIÓN

[EI] EXTRACCIÓN DE INFORMACIÓN

[LDA] ASIGNACIÓN LATENTE DE DIRICHLET (LATENT DIRICHLET ALLOCATION)

[OMS] ORGANIZACIÓN MUNDIAL DE LA SALUD

1. INTRODUCCIÓN

1.1 MOTIVACIÓN

Hoy en día miles de mujeres y niñas alrededor del mundo están sufriendo diferentes formas de violencia y abusos a lo largo de sus vidas. La violencia contra las mujeres, también conocida como violencia de género, ha sido ampliamente reconocida como una grave violación de los derechos humanos y como un importante problema de salud pública que afecta a todos los sectores de la sociedad [1]. Estos abusos dirigidos a mujeres muestran el desequilibrio de poder ejercido desde el individuo más fuerte hacia el individuo más débil, lo que da como resultado una desigualdad de género. Además, esta desigualdad se puede describir como una discriminación con raíces socioculturales mostrando la masculinidad superior a la feminidad. Los actos de violencia contra la mujer (VCM) se han vuelto tan frecuentes que autoridades y organizaciones la han reconocido como un problema de salud pública y una violación de los derechos humanos con consecuencias tanto físicas como mentales [2].

El reconocimiento de la violencia como un problema de salud y de derechos se vio subrayado y reforzado mediante acuerdos y declaraciones en conferencias internacionales durante la década de 1990, como la Conferencia Mundial de Derechos Humanos (Viena, 1993)[1] , la Conferencia Internacional sobre Población y Desarrollo (El Cairo, 1994)[2] y la Cuarta Conferencia Mundial sobre la Mujer (Pekín, 1995) [3]. Por otro lado, en la Declaración para la Eliminación de la Violencia Contra la Mujer de la Asamblea General de las Naciones Unidas se estableció que cualquier acto de violencia que resulte en un daño o sufrimiento físico, sexual o psicológico se considera maltrato, el cual puede generarse en la vida privada o pública. Además, aclara que la mujer debería disfrutar de la protección de los derechos humanos como: derecho a la igualdad, derecho a la libertad, derecho a la seguridad, derecho a no ser objeto de discriminación, derecho a salud tanto física como mental, derecho a condiciones de trabajo equitativas y satisfactorias. Por

[1] https://unesdoc.unesco.org/ark:/48223/pf0000095414_spa
[2] https://www.un.org/es/conferences/population/cairo1994

último, el derecho a no ser sometido a tratos o penas crueles, inhumanos o degradantes [4].

La necesidad de realizar un estudio acerca de la VCM es bastante clara, esta problemática es tan amplia que afecta a la víctima, sus familias y hasta a la comunidad. A parte de las heridas producidas en el cuerpo de la víctima, los actos de violencia pueden ocasionar consecuencias mentales a largo plazo como; depresión, intentos de suicidio o incluso estrés postraumático. En el caso de violencia sexual sus consecuencias pueden ser; la transmisión de enfermedades sexuales, embarazos no deseados o hasta problemas reproductivos. En el caso de violencia contra niñas las consecuencias de las agresiones pueden estar presenta a lo largo de sus vidas [5].

Por lo mencionado anteriormente, en esta investigación se debe realizar un estudio y análisis que permita profundizar en la problemática de la VCM, con el fin, de poder conocer las diferentes características que la involucran, su impacto y su alcance. Según un estudio realizado por la Organización Mundial de la Salud se estima que globalmente 1 de 3 (30 %) de mujeres a nivel mundial han sufrido violencia física o sexual por parte de su pareja o alguien externo en el transcurso de su vida, así mismo, en todo el mundo, casi un tercio (27%) de las mujeres de entre 15 y 49 años que han mantenido una relación declaran haber sido objeto de algún tipo de violencia física y/o sexual por parte de su pareja[3]. Estos datos reflejan claramente que la VCM es un problema que está arraigado en muchas sociedades, además se está volviendo en un problema común. A causa de esta realidad surge la necesidad urgente de prevenirla y erradicarla, con el propósito de poder brindar seguridad e igualdad de género a las mujeres a nivel mundial.

Para apoyar y sumarnos a la lucha contra la prevención y erradicación de la VCM es primordial entender la dimensión de esta problemática en la sociedad, indagando en las diferentes características que la involucran, de manera que se genere una idea más asertiva sobre su impacto y por consiguiente poder concienciar sobre los daños que esta provoca.

Ante el incremento de los casos de la VCM, y el auge del internet como un medio de comunicación donde se comparte información relacionada a la violencia feminicida en

[3] https://www.who.int/news-room/fact-sheets/detail/violence-against-women

diversos sitios web oficiales de la prensa escrita. Es por ello que en la actualidad este tipo de información está siendo utilizada solo como un recurso de lectura y no como un llamado de atención a la sociedad, ya que estas publicaciones pueden ser una rica fuente de información que ayuden al estudio y extracción de conocimiento acerca de la VCM.

Conforme a lo expuesto anteriormente, la creciente ola de casos de VCM a nivel mundial generan la necesidad de luchar contra esta problemática con el objetivo de erradicarla. Es por ello que es imprescindible utilizar Sistemas Inteligentes que apoyen al tratamiento del texto de manera que se pueda desarrollar un estudio que contribuya a extraer conocimientos a fin de generar y profundizar la percepción sobre la VCM. Además, esta información puede ayudar a la creación de nuevas leyes que protejan a las víctimas, así mismo es necesario fortalecer las leyes ya existentes con el propósito de garantizar la igualdad y protección del género femenino.

Para la creación de Sistemas Inteligentes se cuenta con los recursos otorgados por las disciplinas de la Minería de Texto (MT) y Aprendizaje Automático (AA). A la disciplina de la MT se la puede definir como un área orientada a la transformación de texto no estructurado a datos estructurados facilitando su análisis y la extracción de conocimiento. Consiste en un conjunto de técnicas que se orientan al descubrimiento de nuevo conocimiento que era inexistente o poco visible en la colección de documentos (textos) de las que se partían. Si nos enfocamos en las técnicas de la MT es posible obtener información que no ha sido estudiada, ya que se evidencian datos relevantes para determinar los factores comunes en los casos de la violencia contra la mujer, es decir, se logra extraer elementos importantes como lo son patrones sociales que involucran tipos de comportamiento previos a la agresión contra la mujer.

Por otra parte, la disciplina del Aprendizaje Automático es considerada como una rama de la Inteligencia Artificial (IA) que tiene como objetivo la construcción de modelos capaces de adquirir conocimiento y aprender automáticamente en base a un conjunto de datos de entrenamiento y posteriormente con el uso de datos de evaluación poder realizar predicciones relacionadas a la VCM.

Por consiguiente, la disciplina del Aprendizaje Automático se encarga de procesar, analizar y mantener el enfoque principal del texto. Cabe destacar, que una vez se

adquieran datos relacionados a la violencia contra la mujer se fragmentan datos precisos que a la larga pueden contribuir en la prevención del maltrato machista contra la mujer

Gran parte de los sistemas basados en MT se construyen sobre la base del Procesamiento del Lenguaje Natural (PLN) en conjunto con el Aprendizaje Automático. El PLN es una disciplina que combina la lingüística computacional, la informática, la ciencia cognitiva y la inteligencia artificial con el propósito de que los ordenadores puedan procesar o comprender el lenguaje humano (es decir, natural). Todas estas disciplinas son imprescindibles para el procesamiento y análisis de grandes colecciones de texto.

Las ventajas de la aplicación de Minería de Texto y Aprendizaje Automático en la explotación y generación de conocimiento antes oculto en texto referente a casos reales de VCM, sería de apoyo dentro del ámbito social y de la igualdad de género, entre los aportes podemos mencionar: extraer conocimiento que permita profundizar en el problema de la VCM, herramienta para sensibilizar a la sociedad sobre esta problemática, apoyo para la creación de nuevas leyes que ayuden a prevenir la VCM a su vez que promuevan la seguridad e igualdad de las mujeres. En la actualidad existen escasos sistemas que se enfoquen en extraer información relevante de la VCM a partir de texto, y mucho menos en texto escrito en idioma español. Lo que hace que esta investigación sea más compleja, y que el proceso de extracción de conocimiento sea más difícil de conseguir.

1.2 OBJETIVOS

El objetivo principal de esta investigación es aplicar diferentes técnicas de minería de texto y aprendizaje automático sobre los datos textuales (noticias) obtenidos de portales web de la prensa digital para estudiar y analizar su contenido. De este modo, contribuir a la búsqueda y descubrimiento de nueva información que una vez publicada en internet no solo será una fuente de lectura, sino que se convertirán en un medio de información y conocimiento capaz de aportar detalles más profundos acerca del fenómeno de la VCM.

Por lo tanto, de forma esquematizada, señalamos que los objetivos específicos de esta investigación son los siguientes:

- Estudiar y entender la problemática social de la violencia contra la mujer para crear bases de conocimiento previo a la aplicación de las técnicas de minería de texto.

- Explorar y analizar el proceso de Raspado Web para adquirir el conocimiento necesario que nos permita poder aplicarlo en la recolección del conjunto de datos que será estudiado.

- Investigar y evaluar la aplicabilidad las diferentes técnicas de Minería de Texto para definir qué modelos o algoritmos se deben utilizar, con el fin, de obtener óptimos resultados.

- Clasificar noticias en los diferentes tipos de Violencia Contra la Mujer.

- Generar temas latentes a partir de noticias sobre la Violencia Contra la Mujer.

- Extraer reglas de asociación que representen patrones frecuentes y así analizar las diferentes características asociadas a la VCM.

1.3 ESTRUCTURA

Esta memoria está organizada de la siguiente forma:

El capítulo 2 está dedicado al marco teórico de la tesis. En el primer lugar se describirán conceptos sobre la VCM junto a sus datos estadísticos, describiremos los diferentes factores de riesgo que desencadenan actos de VCM, así mismo, los tipos de violencia existentes y reconocidos por la OMS, posteriormente se hablará sobre la violencia que ocurre en el seno familiar y de las medidas que se están tomando para aportar a la erradicación de esta problemática. A continuación, se hablará sobre la recuperación de información aplicando técnicas de raspado web, también se describirá la Minería de Texto, las áreas dentro de la minería de texto, sus etapas de desarrollo y tareas que se pueden realizar aplicando técnicas de minería de texto.

El capítulo 3 está enfocado a realizar una descripción sobre la metodología aplicada en el desarrollo de este trabajo. En la primera sección "Clasificación de Texto" se describirán las tareas realizadas en la recuperación de noticias sobre la VCM además en el apartado del procesamiento del texto se detallarán los procesos aplicados a la colección de documentos para así darle una estructura al texto que facilite su manipulación y estudio, igualmente se escribirán las metodologías aplicadas para el etiquetado del conjunto de entrenamiento. Seguidamente, se detallarán las tareas realizadas en la etapa de transformación de texto. El proceso realizado para balancear las clases a predecir igualmente los algoritmos de clasificación aplicados. Por otro lado, en la tarea de "Generación de Temas Latentes" se detallarán las actividades realizadas para identificar y generar los temas sobre la VCM incluyendo el proceso de creación de sus etiquetas de identificación. Por último, en la tarea de "Extracción de Reglas de Asociación" se describen las actividades de tratamiento de texto y el proceso de generación de las reglas de asociación.

El Capítulo 4 está dedicado a la fase de experimentación y los resultados obtenidos en el proceso de clasificación de textos en los diferentes tipos de VCM, la extracción de temas referentes a la VCM sus distribuciones de temas además de las etiquetas creadas para la identificación. Finalmente, en este capítulo se detallarán las reglas de asociación extraídas (patrones y relaciones existentes) y se analizarán en profundidad los resultados obtenidos para finalizar con las conclusiones de mayor interés.

En el Capítulo 5 se recogerán las conclusiones finales de esta tesis doctoral y las propuestas de mejora o líneas de trabajo que se proponen para el futuro. Para terminar, en el Capítulo 6 se detalla el trabajo científico publicado relacionado a este trabajo doctoral

2 MARCO TEÓRICO

En este capítulo realizaremos inicialmente un análisis detallado sobre la VCM, describiendo conceptos importantes con algunas de sus estimaciones mundiales, cuáles son los factores de riesgo de sufrir violencia, los tipos de violencia reconocidos por la OMS además de la descripción de los mismos, se hablará sobre la violencia que se sufre en el seno familiar y las acciones que se están llevando a cabo para erradicar la VCM. Seguido se detallará el proceso de recuperación de información de páginas web. Después se hablará sobre la minería de texto, las áreas que comprende la minería de texto: procesamiento de lenguaje natural, extracción de información y aprendizaje automático. Finalmente, se describirá el procesamiento y transformación de texto además del modelado y evaluación de resultados.

2.1 VIOLENCIA CONTRA LA MUJER

La Violencia Contra la Mujer (VCM) es un fenómeno universal que persiste en muchos países del mundo, se debe agregar que, a menudo los agresores son bien conocidos por sus víctimas. La VCM tiene un impacto mucho más profundo que el daño directo e inmediato causado a la víctima. Este tipo de violencia tiene consecuencias devastadoras para las mujeres que la sufren y un efecto traumático que puede afectar de por vida a quienes la presencian, especialmente a los hijos de la mujer violentada.

En el pasado muchos gobiernos de diferentes países veían la VCM como un problema social de poca relevancia debido a que creían que afectaba a pocas mujeres, pero al pasar los años un gran número de víctimas lo sufren volviéndose muy evidente el alto impacto que representa tornándose en un tema de importancia a nivel político pero la VCM lastimosamente no en todos los países es reconocida como una forma extrema de desigualdad de género. Se debe agregar que la VCM tiene un alcance negativo tan amplio que el dolor y sufrimiento causado a la víctima trasciende a sus familias.

La Oficina de las Naciones Unidas contra la Droga y el Delito (United Nations Office on Drugs and Crime) en sus estadísticas del año 2021 menciona que el número de mujeres asesinadas por compañeros íntimos o familiares a nivel mundial es de alrededor 45.000 mujeres, significando que más de 5 mujeres son asesinadas por algún familiar cada hora[4]. Estas estadísticas se muestran en la *Figura 1*.

Figura 1. Estadísticas de mujeres asesinadas a nivel mundial.
Fuente: Gender-related killings of women and girls (femicide/feminicide) (UNODC, 2022).

Espinoza Bonifaz [6] señala que muchos casos de violencia ocurren por causa de los roles socioculturales atribuidos a mujeres y hombres donde se cree: "que existe una superioridad masculina". Además, se cree que el "verdadero hombre" debe ser un macho heterosexual con comportamiento agresivo y entre más mujeres posea es mejor, no se puede ni debe seguir tolerando este tipo de criterios que solo contribuyen a la existencia del machismo. Igualmente, prevalece la idea que engañar a una mujer es causa de orgullo,

[4] UNODC, Gender-related killings of women and girls (femicide/feminicide), (2022)

por lo cual resulta normal que el macho pueda conquistar a todas las mujeres posibles, pues su relación con su pareja es la de dueño y protector, acompañado de una superioridad no-sentimental y distante. Otra idea inmersa en la sociedad es que los hombres pueden humillar y golpear a sus mujeres porque "para eso son los maridos", y asumen que el lugar de las mujeres es la casa porque los hombres son de la calle. De esta forma el hombre aparece como el jefe de la casa ante los demás sino perdería su prestigio de macho. Este tipo de conducta solo favorece la presencia de la desigualdad de género. Por otro lado, la existencia de celos en conjunto con una actitud agresiva desencadena actos violentos físicos y peor aún el homicidio de la mujer infiel. El comportamiento violento se espera y se "comprende". Pues la agresividad es otra característica destacada del macho. Cada hombre trata de mostrar su superioridad a los demás reflejando que él es el más masculino, el más fuerte, el más poderoso físicamente. Lo mencionado anteriormente hace que la VCM sea apreciada como normal y en muchos casos aceptada por las victimas haciendo que se vuelva difícil de eliminarla. Debido esto el autor afirma que la estructuración social concede mayor poder a los hombres, volviéndose en uno de los principales pilares de la violencia dirigida contra las mujeres, con el proposito, de mantener a las mujeres en una posición de subordinadas y dominadas por quienes supuestamente son sus superiores.

Ante lo descrito en el párrafo anterior, se puede añadir que un punto de partida para la erradicación de la VCM es transformar este tipo de sociedad liderada por la dominación masculina. Es necesario promover cambios muy profundos a nivel de las estructuras sociales, la cual otorga un mayor poder en los hombres sobre las mujeres. Se debe cambiar la mentalidad de los individuos para poder cambiar sus acciones. En el caso de los hombres para que dejen de ser agresores y en las mujeres para que dejen de ser víctimas, debido a que muchas veces las mujeres ven estos actos de violencia como algo sin importancia y que quizá merecen ser tratadas así.

Debido a su género las mujeres y niñas sufren actos de violencia a lo largo de su vida su estado de vulnerabilidad las convierte en víctimas de violencia y misoginia de forma deliberada y constante [7]. La VCM es un problema tan fuerte que tiene la capacidad de atravesar fronteras, clases y comunidades. En todo el mundo, se calcula que 81.100 mujeres y niñas fueron asesinadas intencionadamente en el año 2021. La violencia contra mujeres y niñas se reconoce como una de las violaciones de los derechos humanos más

frecuentes en el mundo, puede suceder de forma diaria, muchas veces al día, y en todos los rincones del planeta. Sus graves consecuencias físicas, económicas y psicológicas a corto y a largo plazo son un obstáculo que impiden a las mujeres y niñas participar de forma plena e igualitaria en la sociedad.

En un estudio realizado por la Organización Mundial de la Salud (OMS), la Escuela de Higiene y Medicina Tropical de Londres y el Consejo Sudafricano de Investigaciones Médicas se presentaron estimaciones sobre la VCM, los datos de este estudio se pueden observar en la *Figura 2* [8]:

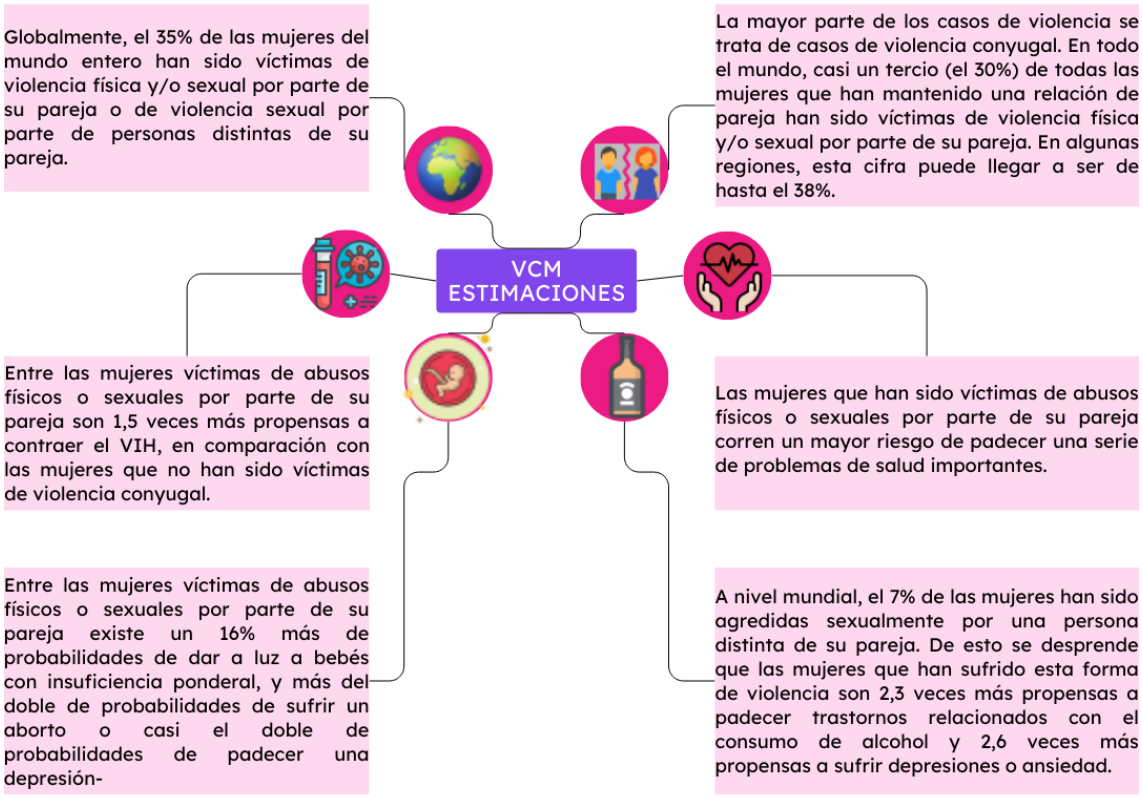

Figura 2. Estimaciones sobre la Violencia Contra la Mujer.

Fuente: Estimaciones mundiales y regionales de la violencia contra la mujer: prevalencia y efectos de la violencia conyugal y de la violencia sexual no conyugal en la salud, (Garcia-Moreno, 2013).

De los datos extraídos de un informe anual de violencia y salud que partió de reportes policiales, estudios de entornos clínicos y organizaciones no gubernamentales, se obtuvieron las siguientes estimaciones, Como ejemplo, en un estudio latinoamericano se calculó que solo alrededor del 5% de las víctimas adultas que han sufrido violencia sexual notificaron el incidente a la policía, pero existen muchos casos donde la víctima no acude a la policía, hay muchas razones lógicas que revelan por qué las mujeres no notifican a

las autoridades correspondientes sobre la violencia sexual sufrida [9]. A continuación, en la *Figura 3* detallamos las posibles razones por las que muchos casos de violencia quedan en la impunidad, se deben analizar para encontrar posibles falencias que se deban mejorar y así contribuir a que las víctimas acudan de forma segura a la policía. Si todos los actos de VCM fueran notificados a las autoridades encargadas de custodiar el bienestar de mujeres y niñas, posiblemente las estadísticas sobre la VCM podrían alcanzar porcentajes más altos de los que se tienen registrados y estudiados.

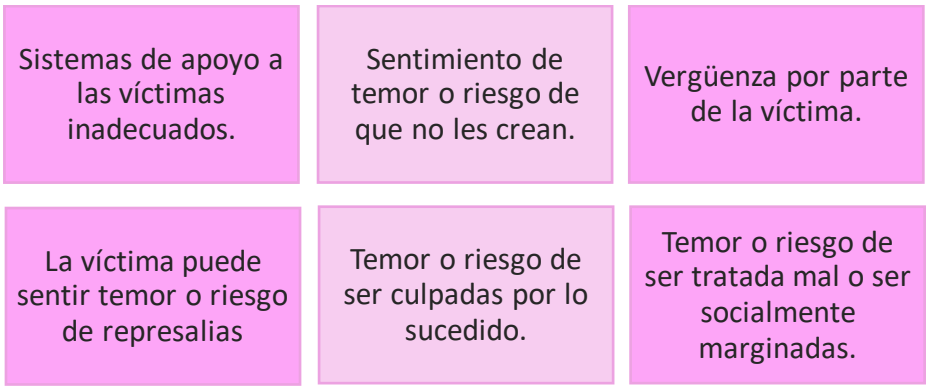

Figura 3. Motivos por los que no se denuncian los casos de abuso a la policía.
Fuente: Comprender y abordar la violencia contra las mujeres, (Gasman et al., 2016).

La VCM con su impacto inhumano ha sensibilizado a organizaciones internacionales, gobiernos y expertos que mediante un arduo esfuerzo han realizado una profunda transformación en la conciencia pública sobre este tema consiguiendo que la violencia contra la mujer, también conocida como violencia de género, se reconozca ampliamente como un grave abuso de los derechos humanos y, cada vez más, también como un importante problema de salud pública que afecta a todos los sectores de la sociedad [8]. Este tipo de violencia quebranta los derechos humanos básicos. De modo que debe eliminarse mediante la voluntad política, la acción legal y civil en todos los sectores de la sociedad, con el fin de, poder brindar una igualdad de género a mujeres y niñas además de asegurarles su activa participación en la sociedad y economía de un país.

2.1.1 FACTORES DE RIESGO DE SUFRIR VIOLENCIA

La VCM tiene asociado algunos factores de riesgo que propician que una mujer sea víctima de violencia. En un estudio enfocado en la violencia se menciona que los factores

que colocan a las mujeres en un estado de vulnerabilidad propensas a sufrir eventos violentos son: las mujeres con bajos niveles de educación, por el contrario, aquellas mujeres que tienen más alto nivel de educación poseen menos posibilidad de sufrir violencia por parte de su pareja. Por otro lado, las mujeres que han tenido experiencias de violencia durante la infancia tienen una mayor posibilidad de presentar violencia o ser victimizadas. Caso contrario, las mujeres que participan en la toma de decisiones en actividades rutinarias en pareja, tienen una menor posibilidad de sufrir violencia de pareja en relación a las que no participan en la toma de decisiones. Otros factores que se pueden relacionar a la violencia contra la mujer son: el consumo excesivo de alcohol, los problemas de personalidad, el bajo nivel educativo, los ingresos económicos precarios, los problemas de desempleo, las experiencias pasadas de violencia, las relaciones con parejas conflictivas y el predominio masculino en la familia [10].

2.1.2 TIPOS DE VIOLENCIA CONTRA LA MUJER

La declaración de la eliminación de la Violencia Contra la Mujer adoptada por la Asamblea General de las Naciones Unidas en 1993, define como violencia de género a cualquier acto que resulte en: un daño físico, sexual o psicológico incluidas las amenazas de tales actos, coacción o privación arbitraria de la libertad, ya sea que ocurra en la vida pública o privada. En consecuencia, las violencias han sido definidas y reconocidas en tres grupos, violencia física, violencia psicológica y violencia sexual[5]. A continuación, mencionaremos los tipos de violencia junto con sus definiciones:

Violencia Física: Refiere a cualquier acto que provoque daño físico como ser abofeteado; que les arrojen algo; ser empujado, acorralado, o tirado por el pelo; golpeado o golpeado con algo que podría causar lesiones; pateado, arrastrado o vencido; asfixiado o quemado intencionalmente; o ser amenazado o asaltado usando una pistola, cuchillo o cualquier otra arma [11]. Como resultado, la violencia física causa daños o heridas que pueden ser de gravedad en el cuerpo de la víctima, dependiendo del nivel de la agresión y los daños recibidos, en muchos casos el resultado de estos actos puede llegar a ser la tortura e incluso la muerte de la víctima.

[5] Fuente: UNWOMEN, violence in five caricom countries: findings from national prevalence surveys on violenceagainst women, (2022)

Violencia Sexual: Consiste en ser forzado a tener relaciones sexuales sin consentimiento mediante amenazas, restricción o infligir dolor; tener relaciones sexuales con una pareja por temor a represalias; o siendo obligados a realizar un acto sexual que la pareja femenina considera humillante o degradante [11].

Gasman et al. [9] añade que este tipo de violencia también puede ocurrir si la persona no está en condiciones de dar su consentimiento, por ejemplo, cuando está ebria, bajo los efectos de un estupefaciente, dormida o mentalmente incapacitada. Además, la misma investigación, indica que la violencia sexual puede causar consecuencias conductuales, sociales y de salud mental a los supervivientes. Conjuntamente, se menciona que las niñas y las mujeres soportan traumatismos y enfermedades como resultado de la violencia y la coacción sexual, no solo porque conforman la gran mayoría de las víctimas sino también porque son vulnerables a sufrir problemas en la salud sexual y reproductiva. En muchos casos se pueden originar embarazos no deseados que desencadenan abortos inseguros o incluso la muerte de la gestante, además estas víctimas tienen un mayor riesgo de contraer inclusive la infección por el VIH. De igual forma, el autor menciona la existencia de factores que provocan que un hombre cometa maltrato sexual a la víctima, estos factores se citan en la *Figura 4*.

Figura 4. Factores que provocan que un hombre cometa abuso sexual.
Fuente: Comprender y abordar la violencia contra las mujeres, (Gasman et al., 2016).

Los factores expuestos en la *Figura 4* reflejan que la violencia sexual es el resultado de múltiples elementos con efectos a nivel individual, relacional, comunitario y social. Por consiguiente, para abordar este tipo de violencia es necesario la cooperación de diversos sectores; de la salud, de la educación, de bienestar social y de justicia penal. Al mismo tiempo se debe asegurar que las víctimas tengan acceso a servicios óptimos y apoyo apropiado.

Violencia Psicológica: Esta violencia refiere al hecho de ser insultado o sentirse mal con uno mismo; ser menospreciado o humillado frente a otros; hacer cosas para asustar o intimidar intencionalmente, o amenazar verbalmente con lastimarla físicamente a ella o a alguien importante para la pareja femenina.

Hernández Ramos et al. [12] mencionan que lastimosamente la violencia psicológica ha sido considerada como un tipo de violencia «invisible» debido a que no se expresa a través de agresiones físicas que puedan servir como evidencia de la violencia sufrida. La violencia psicológica se representa comúnmente con conductas dirigidas a causar daño en la víctima las cuales difícilmente se pueden probar porque al no tratarse de menoscabos o lesiones físicas, no existen huellas visibles en la mujer maltratada. Lamentablemente las huellas o lesiones psíquicas no son fáciles de apreciar o visualizar, por eso, tanto su prueba como su peritación, están sujetas a muchas eventualidades y contradicciones derivadas de la propia «naturaleza interna» de este tipo de lesiones. Por lo tanto, los resultados de la violencia psicológica al no ser «visibles» a la vista humana y al no repercutir de igual manera en todas las víctimas, presentan la «reconocida dificultad de prueba» que beneficia, la mayoría de las veces, a la impunidad del delito. Igualmente, en este estudio se describe que la violencia psicológica es el soporte fundamental en que se sustenta el maltratador para obtener el control total sobre la víctima, socavando su autoestima a través de un progresivo y lento proceso de adaptación a la situación de maltrato, demostrándole a la víctima su poder y autoridad lo cual produce una permanente situación de indefensión aprendida, que propicia que la mujer valore la necesidad de permanecer sumisa e inmóvil frente al agresor, como la única arma para evitar recibir lesiones verbales. Estos actos hirientes infringidos por el agresor pueden causar trastornos psicosomáticos severos, trastornos de personalidad por desestructuración psíquica, agravar enfermedades físicas preexistentes, inducir al consumo de alcohol, drogas o medicamentos no prescritos facultativamente e, incluso, provocar el suicidio en la

víctima. En la *Figura 5* se detallan las principales consecuencias de sufrir cualquier tipo de violencia psicológica.

Figura 5. Consecuencias de la violencia psicológica.

Fuente: El maltrato psicológico. Causas, consecuencias y criterios jurisprudenciales. El problema probatorio, (Hernández Ramos et al., 2014).

A causa del gran impacto psíquico causado por la violencia psicológica y aunque exista dificultad de probar o evidenciar los actos de violentos esta no debe ser sinónimo de impunidad o de poca relevancia, razón por la cual debe evitarse que este tipo de agresores encuentren amparo en ello para eludir, de esta manera, su responsabilidad ante los hechos de agresión psicológica.

2.1.3 VIOLENCIA EN EL SENO FAMILIAR

En relación a los tipos de violencia se debe agregar que otra forma común de Violencia Contra la Mujer es aquella donde las situaciones violentas ocurren dentro del hogar, siendo la pareja o cónyuge el agresor habitual. Por el hecho, de ocurrir dentro del hogar esta violencia también es denominada violencia familiar o violencia doméstica. En investigaciones realizadas se ha logrado demostrar que es más probable que una mujer sea herida, violada o asesinada por un compañero sentimental que por otra persona[6].

[6] Fuente: WHO_violence against women a priority issue.pdf, (1997)

La violencia perpetrada por parte del cónyuge es un problema que ocurre a nivel mundial en todos los estratos socioeconómicos, pero se agrava en los países subdesarrollados debido a las normas socioculturales imperantes [10]. Esta violencia abarca al acto y/u omisión, único o repetitivo, cometido por un miembro de la familia en contra de otro u otros integrantes de esta; estos actos comprenden: el abandono, el maltrato físico, el psicológico, el sexual, el económico [13]. Se ha reconocido que los factores que desencadenan la violencia en la familia pueden ser: **factores sociales;** refiriendo a los patrones dominantes que propician que los niños aprendan desde pequeños que los varones dominan y que la violencia es un medio aceptable en su personalidad, en cambio a las niñas se les enseña a evitar y tolerar las agresiones, **factores culturales;** se crea invisibilidad del abuso, ciertos consensos sociales imponen naturalidad o legitiman el uso de la violencia en la familia, **factores familiares;** experiencias de violencia vividas en la infancia y la juventud las cuales pasan de generación en generación, **factores económicos;** la precariedad del ingreso salarial que no permite tener una estabilidad financiera en el hogar [14]. También podemos citar otros factores como; el **machismo;** donde el hombre coacciona conductas indeseadas causando de forma deliberada dolor, daño emocional, psicológico y físico; **las adicciones;** el consume desmedido de alcohol, o drogas pueden crear un ambiente de constante zozobra y frustración a la familia, por último, **la religión;** existen casos donde los miembros de la familia posen distintas creencias lo que puede originar conflictos con efectos violentos. Es pavoroso que una mujer sufra violencia dentro del seno familiar, un sitio donde debería sentirse segura, en paz y protegida. Sin embargo, el hogar se ha convertido en un sitio violento donde se somete a la familia, estos actos son capaces de afectar la salud física y mental de todos los miembros del hogar, además, de que algunas víctimas lastimosamente son asesinadas por sus agresores.

Arias et al. [10] señala que dentro del hogar el uso de la fuerza física es un método que se suele usar para solucionar problemas intrafamiliares esto se debe al desequilibrio de poder en las parejas o cónyuges, ya sea en forma constante o momentánea, sometiendo de ésta manera a la mujer a daños físico, psicológico incluso económico. Las estadísticas sobre la violencia doméstica indican: que más de 1 de cada 4 mujeres de 15 años o más que han estado alguna vez en pareja sufrirá violencia física y/o sexual por parte de su pareja a lo largo de su vida. Las estimaciones existentes sobre la prevalencia de la

violencia de pareja durante el embarazo oscilan entre el 2 y el 13,5%, y entre el 0,7 y el 55,1% en 126 estudios transversales y de cohortes (en su mayoría clínicos u hospitalarios) de 52 países de todo el mundo. Además. La violencia de pareja puede tener efectos devastadores en las madres, los fetos y los recién nacidos; los datos mundiales sugieren que las mujeres que sufren violencia de pareja durante el embarazo tienen más probabilidades de que se interrumpa el embarazo y corren un mayor riesgo de sufrir abortos espontáneos, bebés con bajo peso al nacer y partos prematuros. La violencia de pareja durante el embarazo también está relacionada con la depresión materna y otros problemas de salud mental perinatales [15].

Patró Hernández and Limiñana Gras [16] señalan las graves consecuencias de la violencia en el hogar, asimismo, el grave riesgo que representa para el bienestar psicológicos de los menores que son testigos de actos de violencia. En muchos casos no solo son testigos de la violencia familiar, también han sido víctimas de ella. En la *Figura 6* se detallan los diversos problemas psicológicos sufridos por niños expuestos a violencia en la familia. Más aún, los autores del estudio describen que los niños pueden presentar un aumento de conductas agresivas y antisociales (conductas externalizantes) además de una creciente conducta de inhibición y miedo (conductas internalizantes) en comparación con los niños que no sufrieron tal exposición, estos niños también pueden presentar una menor competencia social y un menor rendimiento académico que los niños de familias no violentas.

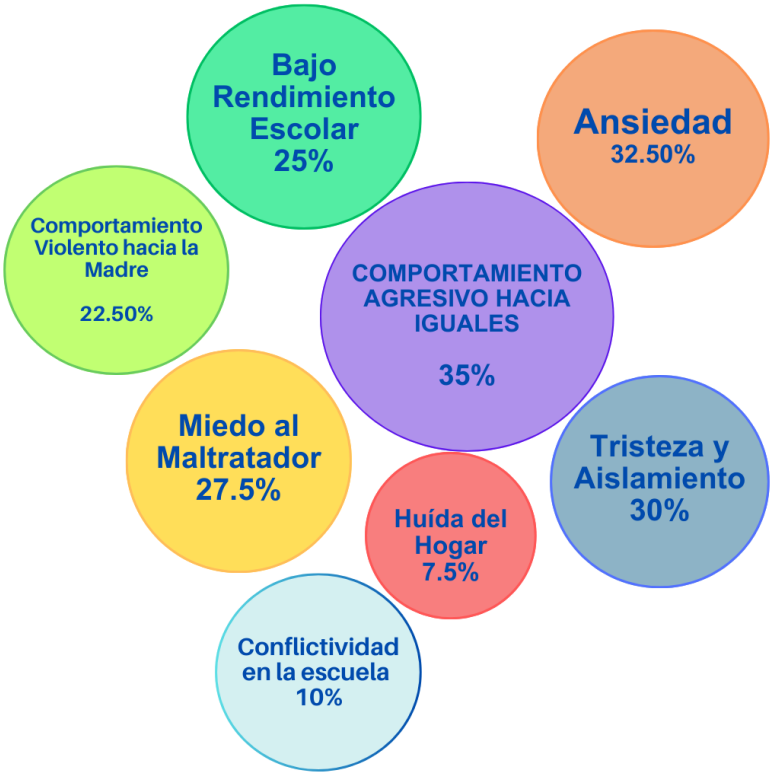

Figura 6. Secuelas de la violencia en los niños.

Fuente: Víctimas de violencia familiar: consecuencias psicológicas en hijos de mujeres maltratadas, (Patró Hernández and Limiñana Gras, 2005).

2.1.4 ERRADICACIÓN DE LA VIOLENCIA CONTRA LA MUJER

Teniendo en cuenta lo mencionado en las secciones anteriores se puede observar que la Violencia Contra la Mujer es un fenómeno que se ha difundido en todo el mundo. Por su parte. La Organización Mundial de la Salud (OMS) estima que un tercio de las mujeres y niñas en todo el mundo sufren violencia en algún momento de su vida. Pero estas cifras son solo la punta del iceberg, ya que este tipo de violencia no suele denunciarse muchas veces debido al estigma y la vergüenza que lo rodean. Como resultado, muchas víctimas esconden su dolor, sufren a solas, y muchos perpetradores permanecen impunes[7].Es por esto, que las mujeres y niñas se han vuelto en individuos vulnerables, frágiles que están constantemente expuestos a sufrir diferentes tipos de violencia a lo largo de sus vidas. Por consiguiente, existen organizaciones, comunidades incluso entidades públicas y

[7] https://ec.europa.eu/commission/presscorner/detail/es/statement_20_2167

gubernamentales que están luchando por erradicar este mal y así poder brindar una vida digna a la mujer. Se reconoce la urgente necesidad aplicar y respetar los derechos y principios relativos a la igualdad, la seguridad, la libertad, la integridad y la dignidad de todos los seres humanos.

La erradicación de la violencia es una de las principales reivindicaciones de las diferentes organizaciones de mujeres que luchan día a día por una vida más segura para ellas y los miembros del seno familiar. Como fruto de este activismo, se logró que tanto en la Plataforma de Acción de Beijing (1995) como en la Convención de la ONU sobre los derechos de las personas con discapacidad (2006), se mencionara y reconociera que la mujer tiene un mayor riesgo de sufrir violencia, así como, la necesidad de implementar ejecutar las medidas y acciones adecuadas para combatirla. En la declaración de la erradicación de la violencia contra la mujer (2013) se reconoció la necesidad de un convenio que permita asegurar que las mujeres disfruten al derecho de protección en pie de igualdad de todos los derechos humanos y libertades fundamentales en las esferas política, económica, social, cultural, civil o de cualquier otra índole[8]. En la lucha constante para combatir la VCM es imprescindible la cooperación de diversos sectores, como los de la salud, de la educación, de bienestar social y de justicia penal. A través de salud pública se desea extender la atención y la seguridad a toda la población poniendo énfasis principalmente en la prevención, velando al mismo tiempo porque las víctimas de violencia cuenten con acceso a servicios y apoyo apropiados. Los datos existentes sobre la prevalencia y los patrones de la VCM pueden ser una herramienta importante para concienciar a los gobiernos y las instancias normativas para que se ocupen del problema y convencerlos de las repercusiones en la salud pública y los costos de la VCM [9].

La violencia contra mujeres e incluso de niñas es una transgresión generalizada de los derechos humanos, un problema sanitario mundial que representa un reto para el desarrollo humano sostenible. Se lo reconoce como un obstáculo importante que impide que las mujeres disfruten de sus derechos humanos y su plena participación en la sociedad y la economía de un país. Por consiguiente, los esfuerzos para eliminar la VCM requiere una comprensión global de la interacción de las normas y actitudes aceptadas sobre los

[8] ONU, Declaración de la eliminación de la violencia contra la mujer (1993)

roles de género y lo que significa ser una mujer o un hombre en la sociedad, así como los factores que dan forma a estas interpretaciones [17].

2.2 RECUPERACION DE NOTICIAS SOBRE LA VCM

Internet es un recurso ampliamente utilizado por la mayoría de las personas en todo el mundo, con más de la mitad de la población accediendo a Internet se transformó en el medio de comunicación más popular incluso que la radio o tv. Es habitualmente utilizado por una persona o grupos de personas para publicar y difundir información sobre diferentes temas de forma libre y rápida, por lo que, se ha convertido en un contenedor masivo de datos. Por otro lado, el intenso uso redes las redes sociales como: facebook, twitter, linkedin, entre otras plataformas conducen a la generación de más información en la web. La mayoría de los datos publicados en línea son accesibles para su manipulación. Estos recursos pueden estar publicados en variados formatos como: páginas web construidas a partir de código HTML; fuentes de datos en formato XML o JSON; datos multimedia (imágenes, archivos de audio o vídeo). Gran parte de estos datos no poseen estructura, por lo que, requieren del uso de técnicas especializadas para acceder a ellos.

A partir de lo anterior surge la importancia del raspado web. El raspado web es una técnica existente considerada eficaz en la recuperación de grandes volúmenes de datos desde Internet, el proceso de recuperación típicamente se la realiza usando palabras claves que hagan referencia al tipo de contenido que se desea obtener. Esta técnica ya ha sido utilizado en múltiples ámbitos como: el comercio minorista, los mercados inmobiliarios y de la vivienda, el mercado de valores, el análisis de las fuentes de los medios sociales, la biomedicina [18], además en muchas áreas de la informática como: Inteligencia de negocios, IA Datos, Big Data, Computación en la Nube y Ciberseguridad. Por otro lado, Singrodia et al. [19] cita que los bioinformáticos construyen raspadores de datos web con el uso de un lenguaje de programación conocido, donde se conceden acceso al sitio con la implementación del lado del cliente del protocolo HTTP, mientras que el análisis sintáctico de las sustancias recuperadas se realiza utilizando funciones de cadena incorporadas como la comparación de expresiones regulares, la tokenización y el recorte.

Existen muchos estudios en los cuales se debe partir desde el proceso de recuperación de información de diferentes fuentes para lograr obtener el conjunto de datos del cual se extraerá conocimiento. Teniendo en cuenta la utilidad del raspado web este se ha convertido en una herramienta popular ampliamente utilizada para recuperar información desde internet en aquellos casos donde los científicos o investigadores no posean un conjunto de datos o repositorio de información para llevar a cabo su estudio y posteriormente obtener nuevo conocimiento a partir de ellos.

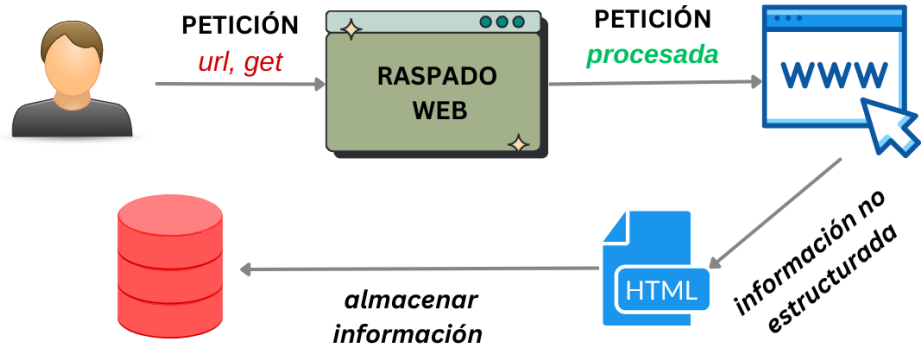

Figura 7. Recuperación de información aplicando técnicas de Raspado Web.
Fuente: Elaboración propia.

En la *Figura 7* se detalla el funcionamiento de las técnicas de raspado web. El proceso empieza cuando el usuario envía una solicitud a través del programa de raspado la misma que puede tener un formato "*url*" con una consulta "*get*" o un segmento de mensaje HTTP que contenga una consulta *post*. Si la solicitud es recibida y procesada con éxito por el sitio web objetivo, el recurso solicitado será recuperado del sitio web y luego enviado de vuelta al programa de Raspado Web. Una vez descargados los datos web, el proceso de extracción continúa para analizarlos, reformatearlos y organizarlos de una forma estructurada [20].

A partir del "Raspado Web" y mediante el procesamiento de la información extraída se producen datos estructurados de información que no poseía formato o estructura alguna. La extracción de los datos es la primera fase de cualquier estudio y desarrollo de ciencia de datos; es el paso en el que los datos se obtienen de fuentes privadas, como registros, informes, o fuentes públicas, como revistas, sitios web y datos abiertos, o mediante la compra de datos. Aplicando Raspado Web se crean procesos automatizados capaces de extraer datos y posteriormente transformarlos en datos estructurados, para luego ser

almacenados en una base de datos externa que permita su análisis o estudio. Su aplicación implica la creación e implementación de dos programas de software: un "crawler" y un "scraper". El crawler se encarga de descargar datos de Internet de forma sistemática; luego, el scraper extrae información importante en bruto de los datos descargados, la codifica y la almacena en una base de datos o un archivo según una estructura y un formato definidos por el usuario [21].

2.3 MINERÍA DE TEXTO

La Minería de Texto (MT) consiste en analizar grandes conjuntos de texto no estructurado con el fin de identificar y aislar temas, palabras claves, conceptos, patrones y otros elementos importantes existentes en el texto. También, se la describe como un proceso de extracción de patrones o conocimientos interesantes y no triviales a partir de documentos de texto no estructurado que busca extraer información significativa de texto escrito en lenguaje natural. Cabe señalar que algunos de los recursos escritos en lenguaje natural son: contenido de páginas web, redes sociales (tweets), noticias, literatura científica, registros y otros tipos de documentos.

La Minería de Texto es una tarea mucho más compleja que la Minería de Datos ya que trabaja con datos no estructurados, a diferencia de la Minería de Datos que trabaja sobre datos estructurados los cuales residen en un campo fijo dentro de un registro o archivo que facilita su manipulación o análisis. Esto implica tratar con texto intrínsecamente desestructurado, difusos e incluso difíciles de tratar algorítmicamente. Los datos no estructurados como el texto suelen referirse a la información que no reside en una base de datos tradicional de filas y columnas. La ventaja de este tipo de minería es que puede funcionar exitosamente en colecciones informativas como mensajes, contenidos y documentos HTML [22].

Con la gran cantidad de datos que se generan diariamente de fuentes como: redes sociales, páginas web y otras aplicaciones centradas en la información, la minería de texto está ganando cada vez más atención y se está convirtiendo en una de las técnicas más prometedoras para el estudio de datos no estructurados. Representa un importante avance tecnológico que tiene la capacidad de enseñar a los computadores el lenguaje natural para

poder realizar tareas que impliquen analizar, comprender e incluso generar texto, una de las ventajas de la MT es que permite analizar grandes volúmenes de texto a gran velocidad.

Se debe agregar que, la minería de texto es un campo interdisciplinar que incorpora diferentes áreas. En la Figura 8 se recogen las áreas más importantes de la minería de texto: Procesamiento de Lenguaje Natural (PLN), Extracción de Información (EI) y Aprendizaje Automático. Estas son la base sobre las que se construyen la mayoría de los sistemas de Minería de Texto y sus técnicas son imprescindibles para llevar a cabo el análisis y procesamiento de grandes colecciones de documentos [23].

Figura 8. Áreas dentro de la minería de texto.
Fuente: Elaboración propia.

En la Figura 8 presenta las áreas que están presente en los sistemas de minería de texto: El procesamiento de lenguaje natural es el que brinda el soporte a los ordenadores para entender el lenguaje humano. La extracción de información es la tarea de extraer automáticamente información estructurada a partir de documentos no estructurados y/o semiestructurados legibles por máquina y otras fuentes representadas electrónicamente. Finalmente, el aprendizaje automático comprende la aplicación de diferentes algoritmos que permitan el estudio del texto para extraer nuevo conocimiento a partir de ellos. En las siguientes secciones se describirá de forma más amplia las áreas de la minería de texto.

2.3.1 PROCESAMIENTO DE LENGUAJE NATURAL (PLN)

Es un subcampo de la inteligencia artificial y la informática que se enfoca en las interacciones y asociaciones entre el lenguaje humano y el informático. También se aplica

en algoritmos de aprendizaje automático para descifrar información en textos y el habla. El PLN trata con datos de texto no estructurados sin formato definido, de aquí surge su grado de complejidad [24]. Su objetivo es explorar cómo pueden utilizarse los ordenadores para comprender y manipular textos en lenguaje natural. El PLN comprende una colección de técnicas computacionales que permiten analizar y entender el lenguaje natural de los humanos. Los investigadores que estudian esta área recopilan conocimientos sobre cómo entienden y utilizan el lenguaje los seres humanos, de esta forma adquieren conocimiento para poder desarrollar herramientas y técnicas adecuadas para que los sistemas informáticos entiendan y manipulen los lenguajes naturales como lo haría un ser humano.

Se puede definir que el lenguaje natural es un conjunto de frases de longitud finita, se construye utilizando un alfabeto finito, o en términos de sintaxis del lenguaje, además de un vocabulario finito de símbolos. Manias et al. [25] describe al PLN como una técnica para convertir el lenguaje humano en formato texto (no estructurado) a datos estructurados (comprensibles por ordenador), pero antes se necesita percibir los datos basándose en la gramática y el contexto. Mediante su aplicación se examina y detecta patrones en los datos y los utiliza para comprender mejor y generar lenguaje natural.

El PLN ha dado lugar a la creación de tecnologías como los asistentes Siri o Alexa, estos son asistentes virtuales capaces reconocer el lenguaje humano (hablado) de los usuarios, se han convertido en dispositivos muy populares a nivel mundial. Las aplicaciones representativas del PLN son el reconocimiento del habla, la comprensión del lenguaje hablado, los sistemas de diálogo, el análisis léxico, el análisis sintáctico, la traducción automática, el grafo de conocimiento, la recuperación de información, traducción automática, grafos de conocimiento, recuperación de información, respuesta a preguntas, la traducción automática, el procesamiento y resumen de textos en lenguaje natural, la recuperación de información multilingüe y translingüística, el reconocimiento de voz, la inteligencia artificial y la traducción experta [26][23].

2.3.2 EXTRACCIÓN DE INFORMACIÓN

En las ciencias de la computación la Extracción de Información tiene como objetivo extraer de forma automática información estructurada, definiendo como estructurada a información categorizada, contextual y semánticamente bien definida, información proveniente de datos no estructurados[9].

Aggarwal [27] describe la Extracción de Información como uno de los problemas clave de la Minería de Texto, que sirve de punto de partida para muchos algoritmos. Una de su aplicabilidad es la extracción de entidades y sus relaciones pudiendo revelar información semántica más significativa en los datos textuales. Con la Extracción de Información el documento escrito en lenguaje natural se convierte en estructurado para luego extraer conocimiento. Posibilita identificar palabras claves y las relaciones dentro del texto buscando secuencias predefinidas en el texto, este proceso se denomina concordancia de patrones. La Extracción de Información logra identificar entidades (personas, lugar, hora, fecha, dirección) del texto, es capaz de proporcionar resultados impresionantes cuando se aplica a un volumen de texto muy grande [28].

La Extracción de información ha sido utilizada en un amplio rango de áreas las cuales serán detalladas en la *Figura 9* seguido de sus respectivas definiciones [22].

[9] https://www.ibm.com/docs/en/db2/9.7?topic=studio-information-extraction

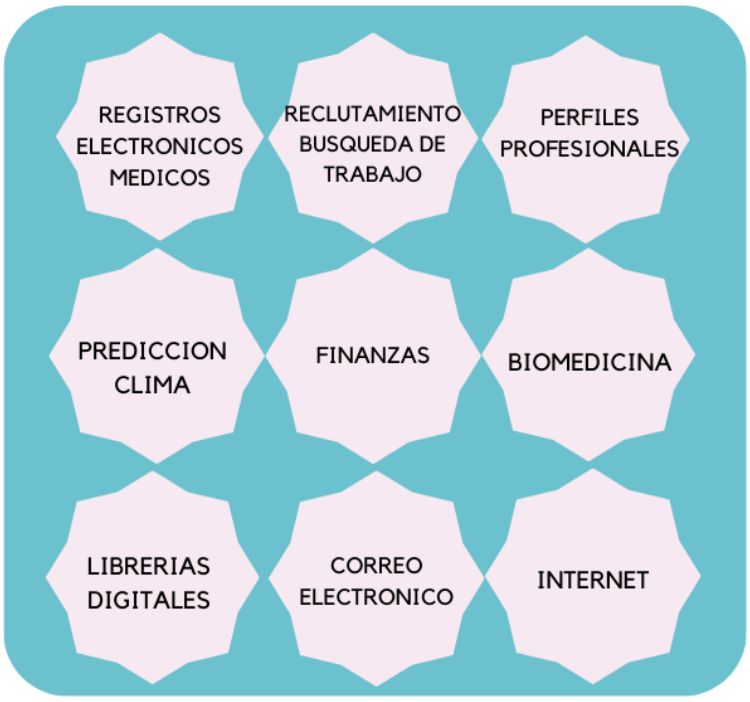

Figura 9. Áreas de aplicación de la extracción de información.

Fuente: Text Mining Challenges and Applications, A Comprehensive Review, (Khan, Khan, et al., 2020).

- ***Biomedicina:*** La Extracción de Información ha sido aplicada sobre investigaciones científicas con el propósito de descubrir información asociada a "genes", "proteínas" u otras entidades biomédicas" específicas [30].

- ***Finanzas:*** A menudo se debe buscar información específica en las noticias para tomar decisiones periódicas, "una empresa financiera puede necesitar conocer todas las absorciones de empresas que se producen durante un determinado periodo de tiempo y los detalles de cada adquisición". Con el reconocimiento de entidades con nombre y la extracción de relaciones se puede descubrir información relevante [27].

- ***Predicción del clima:*** El uso de plataformas de medios sociales en diversas situaciones de emergencia para difundir eventos que suceden en tiempo real muestran la viabilidad de fundar un conjunto automático de servicios destinados a asociar la previsión meteorológica con la detección de acontecimientos y la extracción de información aplicando flujos de medios sociales [31].

- ***Reclutamiento y búsqueda de trabajo:*** Las técnicas de la minería de la información son utilizadas para la contratación electrónica o la búsqueda de

empleo. El número de usuarios en busca de este tipo de servicios crece día a día en todo el mundo [29].

- ***Registros Electrónicos Médicos:*** Los informes clínicos y los resúmenes son una fuente importante que permite revolucionar la investigación relacionada con la salud. Aplicando extracción de información sobre estos datos médicos puede realizarse registros de enfermedades, estudios epidemiológicos, vigilancia de la seguridad de los medicamentos, ensayos clínicos y auditorías sanitarias [32].

- ***Internet:*** Los motores de búsqueda son una herramienta imprescindible para el uso del internet, mediante la extracción de información se puede comprender los comportamientos de búsqueda de los usuarios y así cumplir con sus peticiones [27].

- ***Correos Electrónicos:*** El correo electrónico es una forma sencilla y habitual de comunicarse por texto en el ámbito personal y profesional. Se estima que un usuario puede recibir entre 35 y 45 correos electrónicos al día. Ante este gran flujo de información las aplicaciones de Minería de Texto ofrecen ayuda para realizar las siguientes tareas: filtrado, el enrutamiento y el análisis de correos electrónicos, análisis de grupos de noticias y extracción de información del correo electrónico.

- ***Librerías digitales:*** En la extracción de información en bibliotecas digitales, los "metadatos" que son datos estructurados, ayudan al usuario a descubrir y procesar imágenes y documentos de texto. Los motores de búsqueda utilizan los "metadatos" para recuperar los documentos necesarios con mayor precisión [27].

- ***Perfiles profesionales:*** Un individuo puede tener distintos tipos de información que se relacionan entre sí: perfil personal (página de inicio, afiliación, cargo, retrato, documentos y publicaciones), información de contacto (dirección, teléfono, fax y correo electrónico). Por esto, se aplica la extracción de información para la gestión de información personal [33].

2.3.3 APRENDIZAJE AUTOMÁTICO

La rápida adopción de internet está generando una gran cantidad de datos diariamente. Aquí surge el Aprendizaje Automático (AA) como una herramienta esencial y efectiva para el análisis de datos complejos. Aplicando modelos de AA se puede de forma automática extraer características especiales de grandes conjuntos de datos. Los métodos de aprendizaje automático (por ejemplo, el aprendizaje profundo) se han convertido en una fuerza motriz para revolucionar una amplia gama de sectores, como la sanidad inteligente, la tecnología financiera y los sistemas de vigilancia. Mientras tanto, la privacidad se ha convertido en una gran preocupación en esta era de la inteligencia artificial basada en el aprendizaje automático [34].

Los algoritmos de aprendizaje automático están basados en la Inteligencia Artificial "IA", están transformando la forma en que abordamos tareas del mundo real que son realizadas por humanos. El uso del aprendizaje automático está en ascenso, a través de, la automatización de diversas facetas a nivel social, científico e incluso en negocios [35]. En el aprendizaje automático el ordenador adquirirá conocimientos de los datos de entrada, o bien el propio ordenador debe tener la capacidad de adquirir conocimientos, de resumirlos y mejorarlos continuamente en la práctica [36].

La investigación del aprendizaje automático se lleva a cabo principalmente: (1) para estudiar el mecanismo de aprendizaje humano y el proceso de pensamiento del cerebro humano, (2) para estudiar los mecanismos de aprendizaje de las personas, y (3) para estudiar los métodos de aprendizaje automático y establecer un sistema de aprendizaje para tareas específicas [37]. La investigación del aprendizaje automático se basa en diversas disciplinas, como la ciencia de la información, la ciencia del cerebro, la neuropsicología, la lógica y las matemáticas difusas [38].

2.3.3.1 TIPOS DE APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es un campo de las ciencias de la computación que adapta varios enfoques de aprendizaje que son: el aprendizaje supervisado, no supervisado y

aprendizaje por refuerzo. A continuación, se describen los algoritmos de aprendizaje automático aplicados en este estudio organizados por su tipo de aprendizaje.

A. APRENDIZAJE SUPERVISADO

En este tipo de aprendizaje se proporciona un conjunto de ejemplos o partición de entrenamiento con las salidas correctas (etiquetas a predecir) y, tomando como base el conjunto de entrenamiento, el algoritmo aprende a responder con mayor precisión comparando las etiquetas calculadas en la salida del algoritmo con las etiquetas objetivo pertenecientes a los ejemplos de entrada [39]. Los algoritmos que se centran en el aprendizaje supervisado hacen una búsqueda de una función que asigne los datos de entrada a la salida deseada. Posteriormente, esta función puede aplicarse al nuevo conjunto de datos de entrada para predecir la salida. Uno de los objetivos es minimizar el sesgo y el error de varianza de los resultados predichos. El error de sesgo se debe a suposiciones simplificadoras que realiza el algoritmo de aprendizaje para facilitar el aprendizaje de la función objetivo. Sin embargo, los métodos con un sesgo elevado tienen un menor rendimiento predictivo en problemas que no satisfacen totalmente los supuestos. Citando como ejemplo, un modelo lineal no podrá ofrecer predicciones precisas si los datos subyacentes poseen un comportamiento no lineal. Por otro lado, la varianza muestra en qué medida cambian los resultados del algoritmo de Aprendizaje Automático para un conjunto de entrenamiento diferente. Cabe mencionar, entre más complejidad presente el modelo, más datos de entrenamiento serán necesarios para entrenar el algoritmo y obtener una predicción precisa del modelo. Por otra parte, cuando la dimensión de los datos es alta, el resultado puede depender de una combinación enrevesada de factores de entrada, lo que requiere un elevado número de muestras de datos para detectar las relaciones entre los factores de entrada y los de salida [40]. En la *Figura 10* se puede observar una representación de un modelado mediante aprendizaje automático supervisado.

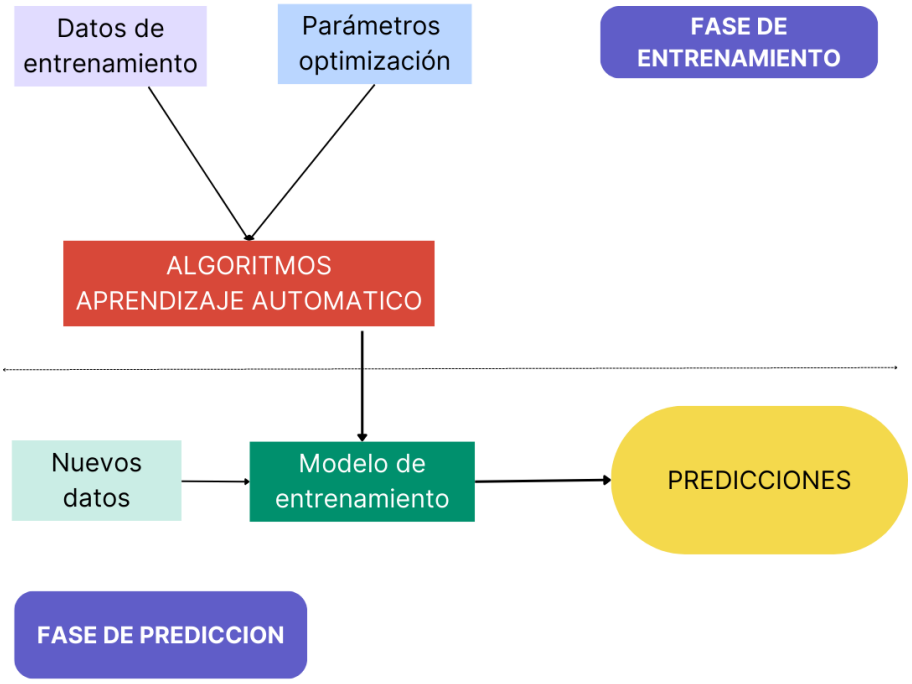

Figura 10. Estructura de aprendizaje supervisado.

Fuente: Elaboración propia.

A continuación, describiremos los algoritmos de aprendizaje automático supervisado aplicados en esta investigación:

A.1 ARBOL DE DESICIÓN

Un árbol de decisión representa elecciones y sus resultados en forma de árbol donde los nodos del gráfico representan un suceso o elección y las aristas del gráfico representan las reglas o condiciones de decisión, una representación de esto se puede ver en la *Figura 11*, cada árbol consta de nodos y ramas, cada nodo representa atributos de un grupo que debe clasificarse y cada rama representa un valor que puede tomar el nodo [41]. Cada uno de estos nodos internos representa una prueba concreta utilizada para clasificar instancias (por ejemplo, "¿El paciente es hombre o mujer?"). Para cada resultado posible de una prueba, hay un nodo hijo [42].

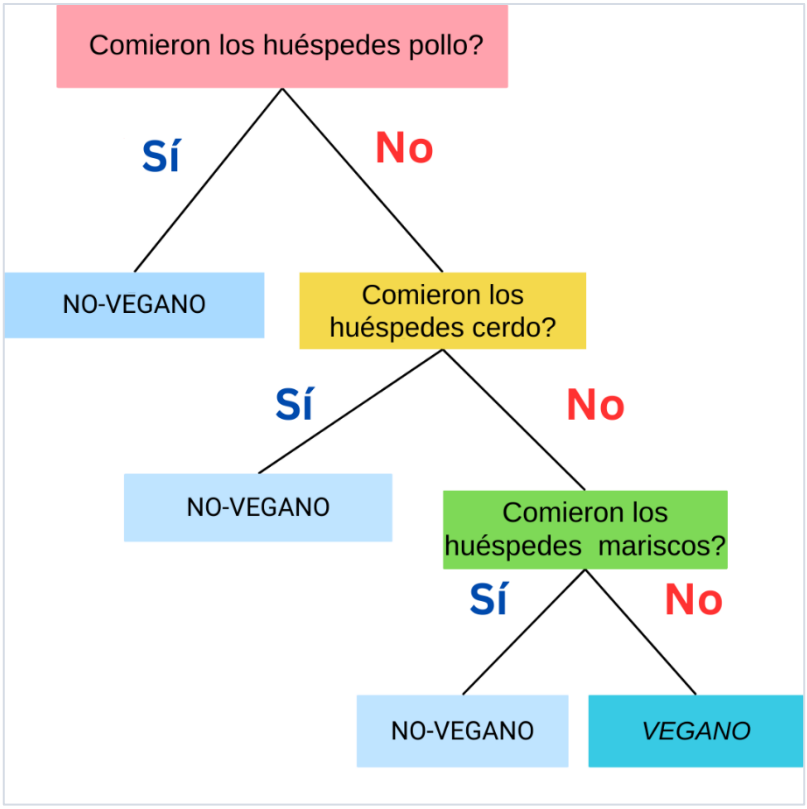

Figura 11. Representación de un árbol decisión.
Fuente: Machine Learning Algorithms - A Review, (Mahesh, 2020).

Un árbol de decisión forma un árbol dirigido con un nodo llamado "raíz" que no posee aristas entrantes, es el padre de todos los nodos y es el nodo superior en la estructura de árbol [43]. En estos árboles un nodo con aristas salientes se denomina nodo interno o de prueba. Todos los demás nodos se denominan hojas (también conocidos como nodos terminales o de decisión). La clasificación aplicando árboles de decisión destaca entre otras herramientas de apoyo a la toma de decisiones por varias ventajas, como su sencillez de comprensión e interpretación, incluso para usuarios no expertos [44].

Los árboles de decisión forman parte de los métodos más potentes aplicados en diversos campos, entre ellos: el aprendizaje automático; el procesamiento de imágenes y la identificación de patrones, se utiliza principalmente con fines de agrupación. Además, son un algoritmo de clasificación utilizado habitualmente en minería de datos debido a su sencillez de análisis y su precisión en múltiples tipos de datos por lo que los árboles de decisión han encontrado muchos campos de aplicación [45].

A.2. REDES BAYESIANAS

Las Redes Bayesianas son un clasificador muy sencillo y eficaz que puede utilizarse tanto para categorías binarias como multiclase en muchos sucesos del mundo real, como la clasificación de documentos o textos, el filtrado de spam, y más [46]. Está basado en la teoría de Bayes que consiste en calcular la frecuencia de ocurrencia de algún suceso en el pasado para estimar la probabilidad futura de que ocurra. En concreto, se utiliza principalmente con fines de agrupación y clasificación en función de la probabilidad condicional de que ocurra algo [41]. La fórmula bayesiana se describe a continuación [47]:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)*P(A)}{P(B)} \qquad (1)$$

En la formula anterior $P(A|B)$ corresponde a la probabilidad de que suceda A dado el suceso B. además $P(A)$ es la ocurrencia de A y $P(B|A)$ es la probabilidad de ocurrencia del evento B cuando ocurre el evento A, por ultimo $P(B)$ es la probabilidad de ocurrencia de B [48]. Caso contrario para saber la probabilidad de que se produzca el suceso B sabiendo que el suceso A ya ha ocurrido. Se debe utilizar la siguiente fórmula [47]:

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(B|A)*P(B)}{P(A)} \qquad (2)$$

En el algoritmo de clasificación bayesiano $P(A|B)$ refiere a la probabilidad de que el documento A sea etiquetado con la clase B [49]. Por lo tanto, si el documento que se desea clasificar está representado por A los resultados de la clasificación estarán representados por B, entonces la formula se puede representar de la siguiente forma [47]:

$$P(documento|clase) = \frac{P(clase|documento)*P(documento)}{P(clase)} \qquad (3)$$

Las redes bayesianas pueden predecir la clase de nuevos sucesos utilizando probabilidades aprendidas a partir del entrenamiento con datos históricos logrando un

buen rendimiento de clasificación [42]. Este algoritmo está dirigido principalmente al sector de la clasificación de textos. En el contexto de la clasificación de textos, la suposición de este algoritmo es que la probabilidad de que cada palabra aparezca en un documento es independiente de la aparición de otra palabra en el mismo documento [50].

En concreto, las redes bayesianas se consideran una poderosa herramienta empleada en problemas de clasificación, que funciona bien con conjuntos de datos desequilibrados y con valores faltantes [51]. Además, han demostrado tener una alta precisión y velocidad.

Existen principalmente dos modelos para las redes bayesianas el *Modelo Multivariante Bernoulli y* el *modelo Multinomial.* El modelo multivariante tiene una representación binaria de las características, mientras que el segundo modelo representa las características con la frecuencia de términos [52]. En aplicaciones reales, los clasificadores redes bayesianas multinomiales suelen funcionar mejor solo requieren una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios para la clasificación. Este clasificador ha mostrado una alta precisión y velocidad cuando se aplicaron a grandes bases de datos [53].

Las ventajas de las redes bayesianas es que utilizan una técnica muy intuitiva, a diferencia de las redes neuronales, no tienen varios parámetros libres que deban ajustarse simplificando considerablemente el proceso de diseño. Dado que el clasificador devuelve probabilidades, es más sencillo aplicar estos resultados a una gran variedad de tareas. Adicionalmente, no requieren grandes cantidades de datos antes de comenzar el aprendizaje y son computacionalmente rápidos a la hora de tomar decisiones [54].

A.3. MÁQUINAS DE SOPORTE VECTORIAL (SVM)

En el aprendizaje automático, las máquinas de soporte vectorial son algoritmos capaces de analizar los datos para la clasificación, de igual forma, son eficaces en la regresión. Es un algoritmo potente y flexible que cuenta con la capacidad incluso de detección de valores atípicos. El algoritmo realiza la clasificación creando un hiperplano lineal de máximo margen que separa las clases. Este margen hace que haya pocas posibilidades de separar los datos de la muestra y, por tanto, hay pocas posibilidades de clasificar

erróneamente nuevas instancias [55]. También, se define a SVM como un clasificador discriminativo basado en encontrar un hiperplano para separar los datos multidimensionales en diferentes clases. Es un clasificador binario, pero pueden ampliarse a clasificador multiclase [56].

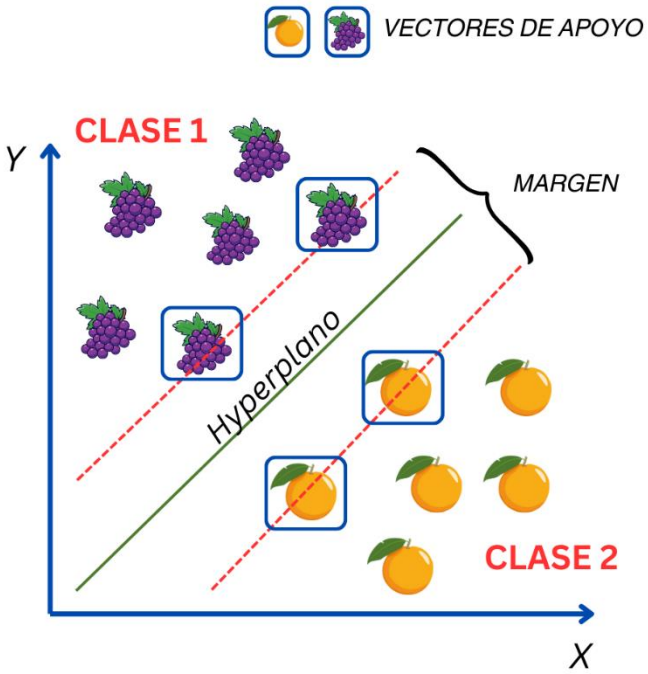

Figura 12. Representación de un modelo SVM.
Fuente: Elaboración propia.

En la *Figura 12* la línea verde es la que mejor separa las clases y es el hiperplano marginal máximo. Los datos dentro de un cuadro en las líneas punteadas son los vectores de soporte y son los factores decisivos en la categorización [52]. SVM también puede modelar límites de decisión no lineales en el espacio de características base aplicando diferente kernel (Se define como kernel a la función matemática que usa el algoritmo con el cual toma los datos de entrada y los transforma a la forma deseada). Entre los kernel existentes tenemos Lineal, Polinómico, Función de base radial (RBF) y Sigmoide[10]. En la *Figura 13* se describen los kernel disponibles para el modelado con SVM. El kernel es la función matemática utilizada para la separación de los datos en diferentes clases, también llamada "núcleo". SVM dibuja márgenes entre las clases de tal manera que la distancia entre el margen y las clases sea máxima y, por tanto, se minimice el error de clasificación [41].

[10] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

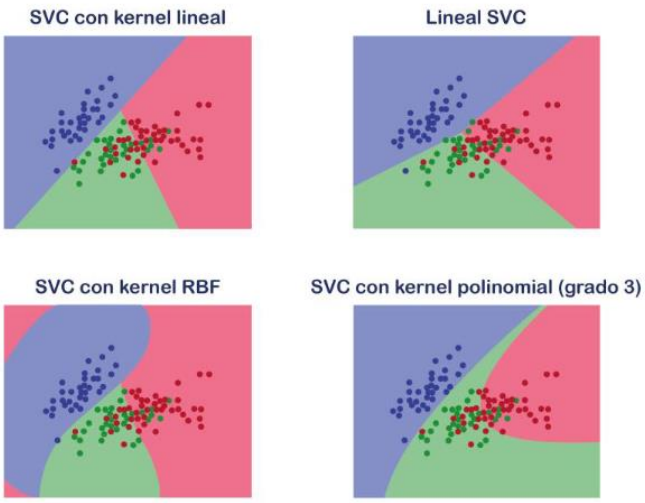

Figura 13. SVM y sus kernel.

Fuente: Support Vector Machines, https://scikit-learn.org/stable/modules/svm.html

SVM con su extraordinaria capacidad de generalización, junto con su solución óptima y su poder de discriminación, ha atraído la atención de la minería de datos, el reconocimiento de patrones y las comunidades de aprendizaje automático. Se ha demostrado que SVM es superior a otros métodos de aprendizaje supervisado debido a sus buenos fundamentos teóricos y a su buena capacidad de generalización, convirtiéndose en uno de los métodos de clasificación más utilizados [57]. Además, ha sido reconocido como uno de los métodos de clasificación de texto más efectivos, siendo capaz de manejar grandes espacios de características y una alta capacidad de generalización [53].

A.4. BOSQUE ALEATORIO (RANDOM FOREST)

El bosque aleatorio o también llamado “Random Forest” es un algoritmo basado en árboles de decisión que fue desarrollado para resolver problemas de clasificación y regresión [58]. Este algoritmo aplica un enfoque que crea un grupo de árboles de decisión con un subconjunto aleatorio de datos. La salida de todos los árboles de decisión en el bosque aleatorio se combina para crear los árboles de decisión finales.

Un clasificador de bosque aleatorio es una técnica de clasificación que se utiliza en el campo del aprendizaje automático y la ciencia de datos además cubre diversas áreas de aplicación. Un bosque aleatorio emplea el "ensamblaje paralelo", que ajusta varios

clasificadores de árboles de decisión en paralelo, en diferentes submuestras de conjuntos de datos y mediante la votación por mayoría o los promedios obtiene el resultado final o la clase a predecir así se minimiza el problema del sobreajuste y aumenta la precisión de la predicción y el control [46]. En la *Figura 14* se puede observar una representación de una estructura de bosques aleatorio.

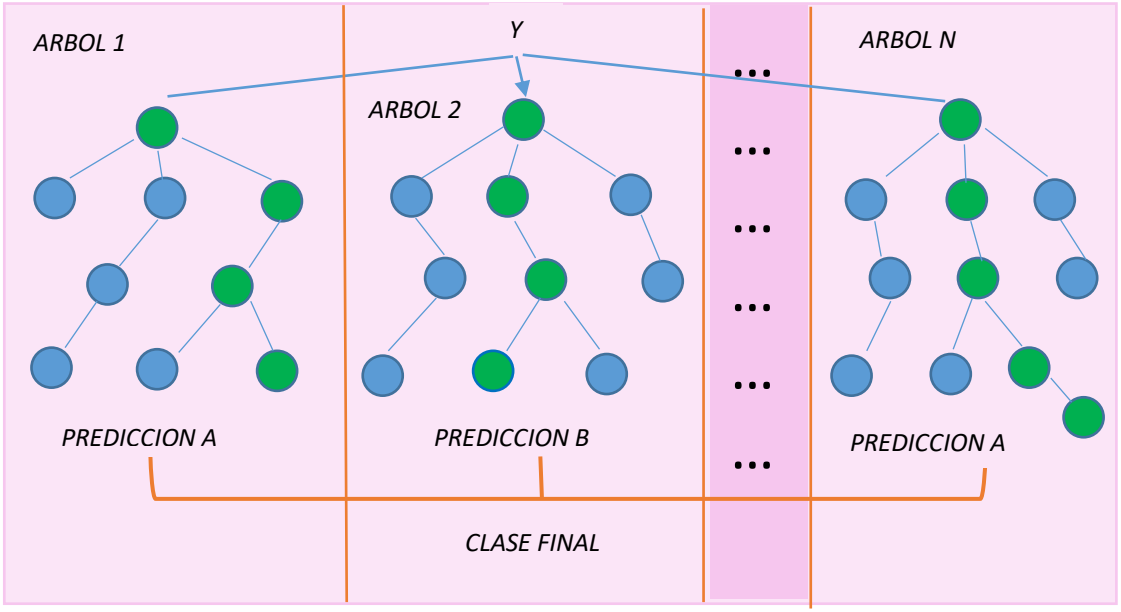

Figura 14. Representación de un bosque aleatorio.
Fuente: Elaboración propia.

Como ya se dijo antes los bosques aleatorios pueden utilizarse tanto en la clasificación, como en la regresión. Entre sus características tenemos: las variables predictoras pueden ser categóricas o continuas, pudiendo ser relativamente rápidos de entrenar y predecir también se pueden utilizar directamente para problemas de alta dimensión permitiendo la detección de valores atípicos y la imputación de valores perdidos [59]. Además, este algoritmo es muy conocido por tratar el desequilibrio de clases especialmente en grandes conjuntos de datos. La arquitectura que posee este clasificador lo vuelven más rápido comparado con otros clasificadores de última generación [60].

Desde su aparición en 2001 por Breiman, el algoritmo se ha vuelto popular en: la clasificación, predicción, el estudio de la importancia de las variables, la selección de variables y la detección de valores atípicos. Siendo ampliamente adoptado y aplicado como clasificador en diferentes áreas: bioinformática [61], visión por ordenador [62], clasificación de imágenes [63]. Este algoritmo ganó popularidad en la clasificación

debido a su proceso de toma de decisiones claro y comprensible, además de sus excelentes resultados de clasificación [64], es un algoritmo prometedor que ha sido incluso utilizado para detectar fraudes a través de datos textuales [65].

Como se ha mencionado antes el algoritmo bosque aleatorio posee un amplio campo de aplicabilidad, esto se puede atribuir principalmente a su capacidad de manejar eficientemente las tareas de clasificación. Además, de ser robusto frente a los valores atípicos y el ruido. Al mismo tiempo es un algoritmo computacionalmente más ligero de procesar en comparación con otros.

A.5. XGBOOST

XGBoost es un algoritmo supervisado ampliamente usado por científicos de datos en el campo del aprendizaje automático. Utiliza procesamiento en paralelo, poda de árboles y manejo de valores perdidos y regularización para evitar el sobreajuste o sesgo del modelo [66]. Este algoritmo se caracteriza por su rapidez de cálculo y su buen rendimiento. Gracias a su escalabilidad ha tenido éxito en una amplia variedad de escenarios, cabe estacar que, funciona más de diez veces más rápido que las soluciones existentes en una sola máquina y puede escalarse a miles de millones de usuarios [67].

El algoritmo XGBoost función de la siguiente forma [68]:

a) Se obtiene un árbol inicial F_0 para predecir la variable objetivo "y", el resultado se asocia con un residual (y - F_0).
b) Se obtiene un nuevo árbol h_1 que ajusta al error del paso previo.
c) Los resultados de F_0 y h_1 se combinan para obtener el árbol F_1, donde el error cuadrático medio de F_1 será menor que el de F_0:

$$F_1\,(X) < -F_0\,(X) + h_1\,(X) \qquad (4)$$

d) Este proceso se sigue iterativamente hasta que el error es minimizado lo más posible de la siguiente forma:

$$F_m\,(X) < -F_{m-1}\,(X) + h_m\,(X) \qquad (5)$$

XGBoost consiste en un ensamblado secuencial de árboles de decisión (este ensamblado se conoce como CART, acrónimo de "Classification and Regression Trees"). Los árboles se agregan en secuencia a fin de aprender del resultado de los árboles anteriores y así corregir el error producido por los mismos, hasta que ya no se pueda corregir más dicho error, Este proceso se conoce como "gradiente descendente" [68]. Entre las ventajas de XGBoost tenemos que es un algoritmo muy preciso y puede ser utilizado tanto para procesos de clasificación como de regresión. Otras de sus ventajas es que requiere un mínimo de ingeniería de características, como la normalización de datos y el escalado de características, ya que el algoritmo puede manejar estas situaciones. Además, es considerado más rápido que la mayoría de los algoritmos de Machine Learning debido a que puede manejar grandes conjuntos de datos y no es propenso al sobreajuste [69]. El buen rendimiento de este algoritmo ha provocado que sea aplicado en diferentes disciplinas como: identificación de huellas digitales, seguridad vial y análisis de mercados financieros y muchos más.

A.6. RED NEURONAL ARTIFICIAL

En el campo de la inteligencia artificial las redes neuronales, se han convertido en un tema candente y apasionante de las tecnologías de la información y la comunicación. Son un tipo de modelo de aprendizaje automático que ha llegado a ser relativamente competitivo frente a los modelos estadísticos y de regresión convencionales en cuanto a utilidad. Las redes neuronales son modelos digitalizados de un cerebro humano, se puede decir que son programas informáticos diseñados para simular la forma en que el cerebro humano procesa la información. Las redes neuronales tienen la capacidad de aprender (o se entrenan) a través de la experiencia obtenida de los ejemplos de aprendizaje (o del conjunto de entrenamiento) [70].

Una red neuronal artificial es un modelo computacional de inspiración biológica formado por cientos de unidades individuales, neuronas artificiales, conectadas con coeficientes (pesos) que constituyen la estructura neuronal [70]. Son estructuras complejas compuestas de neuronas. Una estructura típica de una red neuronal se muestra en la *Figura 15* inicialmente la red neuronal recibe un conjunto de datos, los evalúa e inicia un proceso de entrenamiento para ajustar los pesos de las interconexiones entre neuronas. Su

entrenamiento será supervisado si se conoce la salida, y no supervisado en caso contrario. Hay cuatro variables básicas que caracterizan a una RNA: la topología; el método de entrenamiento; el tipo de asociación entre los datos de entrada y de salida, y la presentación de la información [71].

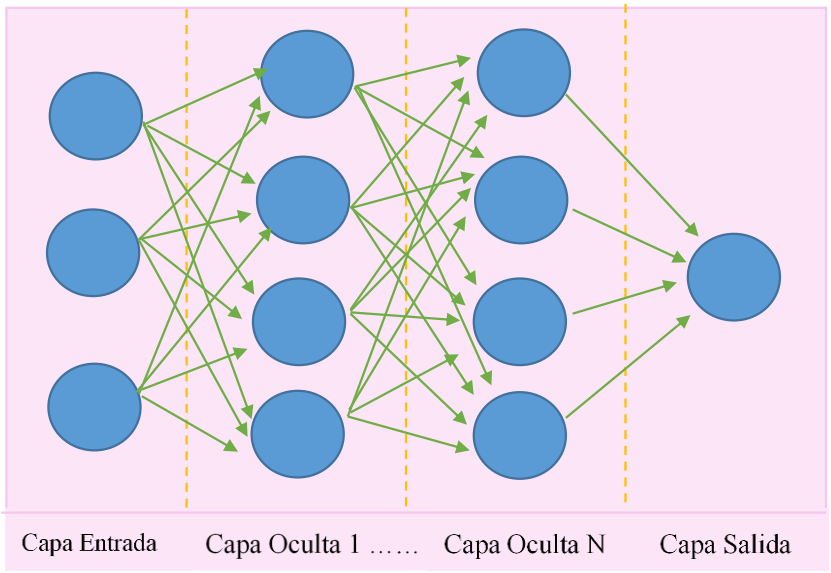

Figura 15. Estructura de una red neuronal.

Fuente: Elaboración propia.

La neurona artificial son un componente importante de las redes neuronales fueron diseñadas para simular la función de la neurona biológica. La neurona recibe señales, llamadas entradas, estas se multiplican por los pesos de conexión (ajustados) se suman (combinan) y luego pasan por la función de transferencia para producir la salida de esa neurona, una de las funciones de transferencia más utilizada es la función sigmoidea. Posteriormente, la función de activación es la suma ponderada de las entradas de la neurona, en la *Figura 16* se puede observar la estructura de una neurona artificial.

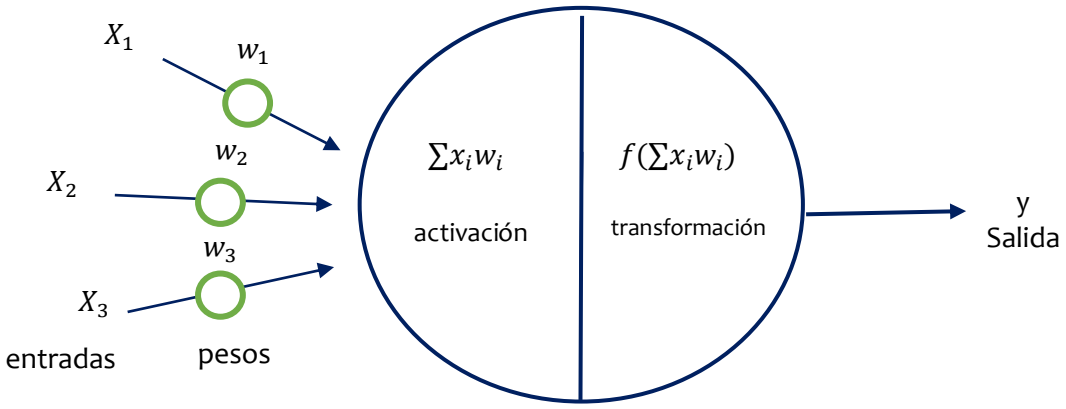

Figura 16. Modelo de una neurona artificial.

Fuente: Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research, (Agatonovic-Kustrin and Beresford, 2000).

Un modelo de red neuronal es tan sencillo y natural que puede manejar problemas muy complejos de la vida real de forma paralela y distributiva como una red neuronal biológica [72]. Por lo que se están convirtiendo en un modelo popular y útil para la clasificación, agrupación, reconocimiento de patrones y predicción en varias disciplinas. Debido al gran potencial y la alta velocidad de procesamiento que ofrecen las redes neuronales han logrado aumentar el interés de investigar en este campo. Además, estos modelos también pueden aplicarse para el reconocimiento de imágenes y el procesamiento del lenguaje natural. En la actualidad se utilizan sobre todo para la aproximación de funciones universales en paradigmas numéricos debido a sus excelentes propiedades de autoaprendizaje, adaptabilidad y tolerancia a fallos [73]. Otros campos de aplicación son la medicina, robótica, química, análisis geoespacial [74] [71]. En la *Figura 17* se describen las ventajas de aplicar redes neuronales.

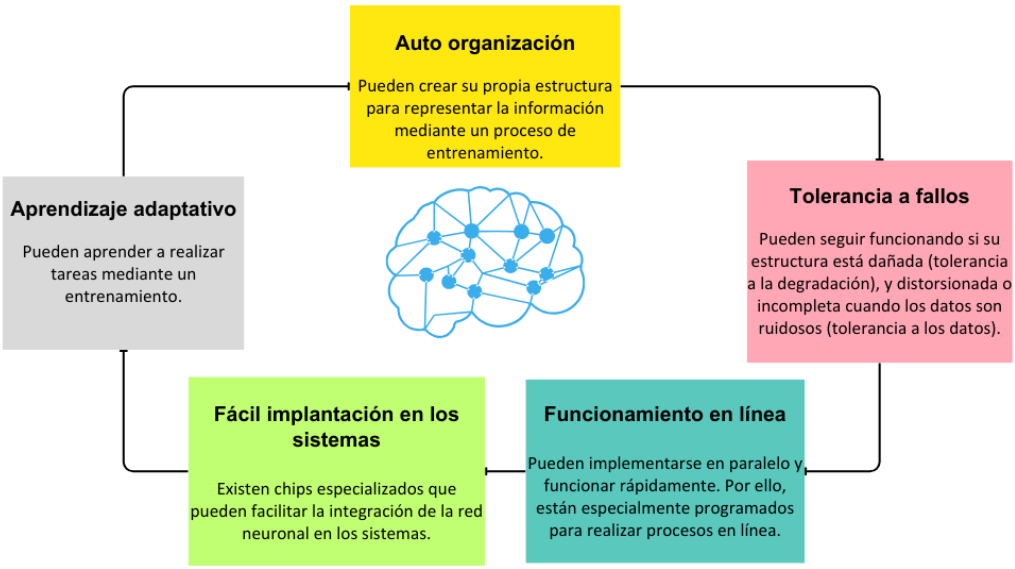

Figura 17. Ventajas de una red neuronal.

Fuente: State-of-the-art in artificial neural network applications: A survey, (Abiodun et al., 2018)

En conclusión, se puede decir que las redes neuronales son modelos simplificados del sistema neuronal biológico y los elementos fundamentales de procesamiento son las neuronas artificiales. Al igual que una neurona natural del cerebro humano, puede recibir entradas, procesarlas y producir la salida correspondiente [75].

B. APRENDIZAJE NO SUPERVISADO

En el aprendizaje no supervisado no hay un conjunto etiquetado previamente. En este tipo de modelado los datos sólo están disponibles en forma de entrada y no hay una variable de salida correspondiente [76]. El aprendizaje no supervisado se considera un enfoque estadístico del aprendizaje y por lo tanto se centra en encontrar estructuras ocultas en datos no etiquetados. Es bastante útil en tareas de descripción, ya que su objetivo es encontrar relaciones en una estructura de datos sin tener un resultado medido. El enfoque de un modelo de aprendizaje no supervisado reside en identificar dimensiones, componentes, conglomerados o trayectorias subyacentes dentro de una estructura de datos [77]. Un caso muy común del aprendizaje no supervisado es la agrupación, aquí se agrupan los datos basándose en similitudes y patrones ocultos. También, se utiliza para problemas como la reducción de la dimensionalidad, en los que se extraen las características clave o más importantes de los datos [40]. Como vemos en la *Figura 18* un

ejemplo de aprendizaje no supervisado son las tareas de agrupación, además incluye las tareas de estimación de la densidad, aprendizaje de características, reducción de la dimensionalidad, búsqueda de reglas de asociación, detección de anomalías, etc. [46].

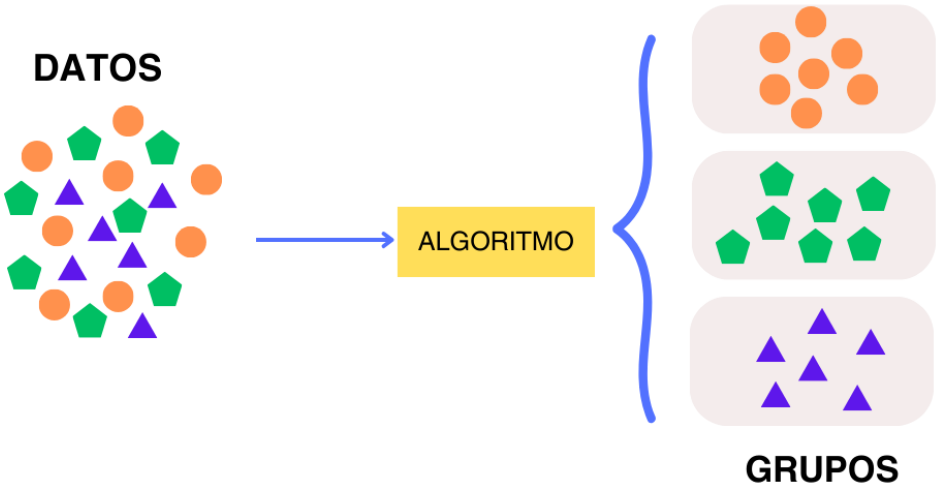

Figura 18. Representación del aprendizaje no supervisado.
Fuente: Elaboración propia.

A continuación, describiremos los algoritmos de aprendizaje automático no supervisado aplicados en esta investigación:

B.1. LATENT DIRICHLET ALLOCATION (LDA)

En la vida cotidiana se generan datos constantemente. La mayor parte de estos datos están creados en un formato no estructurado. Por tanto, generan problemas al intentar extraer información relevante de ellos. Como una forma de resolver este problema surgieron las técnicas de Modelado de Temas. Dichas técnicas usan modelado matemático y estadístico para extraer temas latentes de una colección de documentos. Una de las técnicas ampliamente utilizada es "Latent Dirichlet allocation" (LDA) [78]. LDA se enfoca en representar los documentos como mezclas aleatorias sobre temas latentes, en el cual, los temas generados se caracterizan por una distribución sobre palabras. Las palabras con mayores probabilidades en cada tema suelen dar una buena idea de cuál es el tema.

LDA es un recurso prometedor utilizado en el procesamiento de lenguaje natural para revelar estructuras semánticas denominadas temas [78][79]. Una de sus características es que no requiere un conjunto de documentos previamente etiquetados; por lo tanto, es un

algoritmo no supervisado capaz de identificar temas ocultos dentro de una colección de documentos. LDA es capaz de identificar palabras similares en una colección de N documentos y los agrupa en un número determinado de temas en función de su similitud. A continuación, se detallarán los pasos seguidos por el algoritmo LDA [80]:

a) Considera que hay *k* temas en todos los documentos de un conjunto de datos.
b) Difunde estos m informes entre *k* temas asignando cada palabra a los temas.
c) Cada palabra *w* del informe *m* se asigna al tema correcto.
d) La Asignación probabilística de una palabra *w* a un tema depende de dos cosas:
    i. Qué temas se tratan en el informe *m*
    ii. Cuántas veces se ha asignado a la palabra *w* a un tema concreto en todos los documentos.
e) Repite este procedimiento varias veces para cada registro.

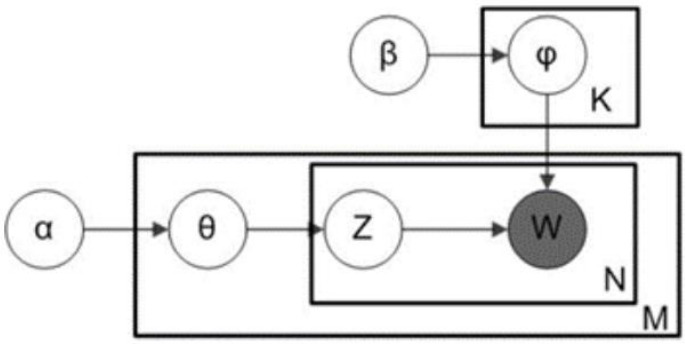

Figura 19. Diagrama de un modelo LDA.
Fuente: Document Representations to Improve Topic Modelling, (Poojitha and Menon, 2020).

En la *Figura 19* los parámetros de la figura representan lo siguiente:

- *N:* Número total de documentos de un conjunto de datos.
- *K:* Numero de temas en un conjunto de datos.
- *N*: Número de palabras que se presentan en el documento
- *M:* Número de documentos contenidos en un conjunto de datos completo.
- *α:* Distribuciones temáticas por documento
- *ß:* Distribución de palabras por tema
- *θ:* Distribución de tema por documento m
- *φ:* Distribución de palabras por tema *k*

- *Z:* Tema para la n-ésima palabra en el m-ésimo documento,
- *w:* Palabra específica en un documento *d*.

En el proceso de aplicación de LDA los datos requeridos son la colección de documentos (artículos sin etiquetar) y el valor del número de temas a extraer [35]. Como resultado, LDA asocia cada tema con un conjunto de términos y proporciona una combinación de los temas generados para cada documento. Cada tema representa los términos clave relacionados con él, es así que, se construye la conexión entre los temas y los términos clave.

LDA se puede utilizar para establecer una relación entre el corpus recopilado y los resultados que se representan estadística y gráficamente. Mediante la aplicación de LDA se puede reconocer patrones ocultos de formación de palabras a través de la distribución de las palabras en los documentos [81], [82]. Cabe destacar que LDA también es de utilidad para resumir, agrupar, vincular y procesar datos muy grandes [83].

Gracias a su popularidad y utilidad se ha convertido en una de las técnicas de modelado más utilizada por diversos investigadores. Además, en el campo de la minería de texto LDA constituye uno de los algoritmos más extendidos considerándose un hito en el panorama del modelado de temas. Es implementado para determinar patrones o tendencias en diferentes campos. Sharma et al. [84] menciona que LDA se aplicado para el análisis de los datos de temas candentes o tendencias publicadas en Twitter, el análisis de reseñas de Internet, la agricultura, la ingeniería de software, el medio ambiente, el aprendizaje profundo, la medicina y muchos campos más.

B.2. APRIORI

Apriori es un algoritmo de la minería de reglas de asociación que se enfoca en encontrar relaciones estrechas entre elementos en grandes conjuntos de datos. Mediante el análisis de los datos y aplicando estos algoritmos, la correlación entre los datos puede ser generados. Se puede decir que la minería de reglas de asociación es utilizada habitualmente para determinar patrones ocultos y relaciones entre las variables de los conjuntos de datos y las dependencias entre estas variables. Los algoritmos de minería de

reglas de asociación se han aplicado con éxito en diversos campos, como la industria sanitaria, el análisis de cestas de mercado y los sistemas de recomendación [85].

Apriori es uno de los algoritmos de reglas de asociación más conocido, es el más popular y ampliamente utilizado. Se lo considera una gran mejora en la historia de la minería de reglas de asociación. Este algoritmo encuentra patrones frecuentes, asociaciones, correlaciones o estructuras informales entre conjuntos de elementos u objetos en bases de datos transaccionales y otros repositorios de información [86]. Así mismo, Apriori se utiliza para encontrar la asociación de palabras clave en las transacciones diseñadas [87].

Este algoritmo es considerado una herramienta importante en el proceso de búsqueda de reglas de asociación sobre un conjunto de datos, los parámetros de entrada son: los conjuntos de ítems, y los umbrales de soporte (cantidad total de transacciones que contienen ambos elementos X e Y) y confianza (proporción de transacciones que contienen el elemento X y también el elemento Y), estos parámetros son imprescindibles, forman parte de las configuraciones para hacer el análisis y descubrimiento de asociaciones. El umbral de soporte se utiliza para generar conjuntos frecuentes y el umbral de confianza se utiliza para generar reglas de asociación a partir de cada conjunto frecuente. Cada regla de asociación está constituida por su ítem condicional y sus conjuntos causales; se interpreta como la afirmación "si el ítem condicional está dado, los conjuntos causales suceden o están dados". El algoritmo Apriori realiza las siguientes tres fases: generación de conjuntos frecuentes, generación de reglas candidatas y selección de reglas. Un conjunto de ítems es frecuente si su soporte está por encima de un umbral de soporte mínimo (α min). Una regla es fuerte cuando su confianza está por encima de un umbral mínimo de confianza [85].

Apriori también se considera el modelo de reglas de asociación más influyente, su aplicación consta de dos pasos: En primer lugar, hay que generar los conjuntos de elementos frecuentes cuyo soporte sea superior al soporte mínimo (minsup) y a la confianza mínima (minconf), se podan todos los subconjuntos de elementos k infrecuentes y, a continuación, se utilizan los conjuntos de elementos frecuentes para producir reglas de asociación [88].

C. APRENDIZAJE POR REFUERZO

El aprendizaje por refuerzo se aplica cuando la tarea en cuestión es tomar una secuencia de decisiones para obtener una recompensa final. Durante el proceso de aprendizaje, un agente artificial recibe recompensas o penalizaciones por las acciones que realiza. Su objetivo es maximizar la recompensa total [76]. En este tipo de aprendizaje los agentes pueden aprender de forma autónoma sin supervisión. La única fuente de conocimiento es la retroalimentación que el agente recibe de su entorno tras ejecutar una acción. Existen dos características importantes en el aprendizaje por refuerzo: ensayo-error y recompensa diferida. La recompensa se define como la señal de retroalimentación que un agente recibe del entorno después de ejecutar cada acción. Estas pueden ser cantidades positivas o negativas, que indican lo buena o mala que es una acción. El objetivo del agente es maximizar estas recompensas explotando el sistema [89].

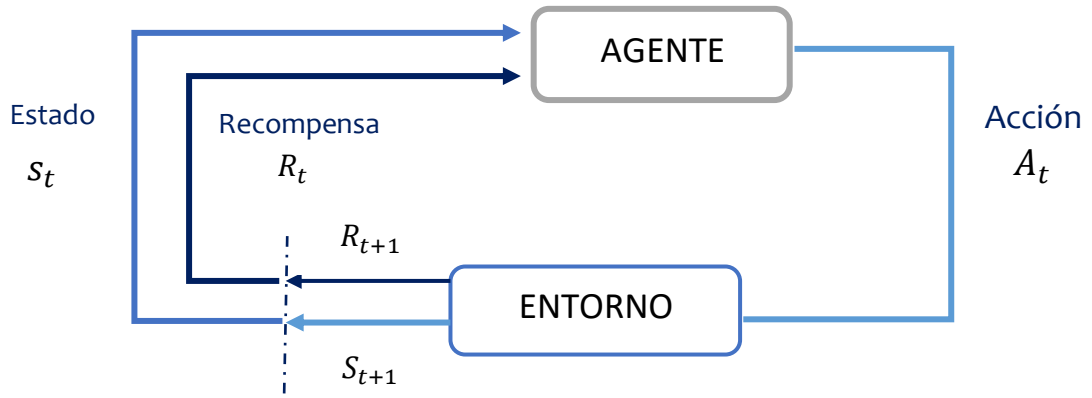

Figura 20. Representación del aprendizaje por refuerzo.
Fuente: Machine Learning Algorithms - A Review, (Mahesh, 2020).

El funcionamiento del aprendizaje por refuerzo se representa en la *Figura 20*. El proceso inicia cuando el agente interacciona con el entorno y realiza una acción. Seguidamente, esta acción afecta al estado del agente en el entorno. Posteriormente, la recompensa se limita a una señal indicando la conveniencia o no de las acciones llevadas a cabo y finalmente con esta señal buscamos mejorar el comportamiento posterior [76]. Como ejemplos del aprendizaje por refuerzo podemos citar a los agentes para juegos de ordenador o para realizar tareas de robótica con un objetivo final.

D. CAMPOS DE APLICACIÓN DEL APRENDIZAJE AUTOMATICO

El aprendizaje automático como parte de las ciencias de la computación y de la inteligencia artificial (IA) proporciona a los ordenadores la capacidad de pensar y aprender sin ser programados explícitamente [90]. Está siendo estudiado y aplicado en muchas disciplinas algunas de ellas las citaremos a continuación [39], [91]:

Visión por computador: Es un campo científico interdisciplinar que hace uso de ordenadores para obtener una comprensión detallada de los datos visuales, un enfoque similar al de los sistemas humanos. Pertenece a la inteligencia artificial y sus aplicaciones han abordado gran variedad de problemas; como la detección de objetos, el reconocimiento facial, el reconocimiento de acciones y actividades y la estimación de la pose humana [92].

Predicción: La predicción cubre los dominios de la clasificación, análisis y recomendación. De esto se desprende: la clasificación de texto, clasificación de documentos, análisis de análisis de imágenes, diagnóstico médico, predicción de detección de intrusiones y predicción de ataques de denegación de servicio. Todas estas tareas se han abordado con éxito utilizando aprendizaje automático [91].

Análisis Semántico, Procesamiento del Lenguaje Natural y Recuperación de Información: El análisis semántico es el proceso de relacionar estructuras sintácticas de párrafos, frases, palabras con el nivel de escritura en su conjunto [91]. El procesamiento del lenguaje natural consiste en programar ordenadores para que puedan leer, interpretar y deducir el significado del lenguaje humano [93]. Podemos decir que la Recuperación de Información (RI) es el procedimiento de representar, almacenar y buscar en un documento o en una colección de ellos con el objetivo de extraer conocimiento y así poder encontrar resultados relevantes que satisfagan las necesidades del usuario ante una consulta de éste [94].

Reconocimiento de voz: En la actualidad, los sistemas de reconocimiento de voz más sofisticados emplean algoritmos de aprendizaje automático. Estos sistemas pueden

aprender sonidos y palabras específicas del hablante a partir de las señales emitidas del habla. Mediante la aplicación de varias técnicas de aprendizaje automático sobre grabaciones de audio en entornos reales se logra detectar impositores y determinar que algortimos o combinación de ellos funciona mejor para el reconocimiento y la clasificación de hablantes [95].

Filtrado de correo electrónico: El aprendizaje automático es de utilidad para filtrar los mensajes de spam. Un modelo basado en aprendizaje automático tiene la capacidad de memorizar todos los correos electrónicos clasificados por el usuario como spam. Posteriormente, cuando se reciba un nuevo correo en la bandeja de entrada, el modelo buscará, comparará y se basará en los correos spam anteriores. Si el correo recibido coincide con alguno de los que ya fueron clasificados como spam este se marcará como spam o correo no deseado caso contrario se trasladará a la bandeja de entrada del usuario. Además con la gran cantidad de mensajes que se reciben a diario el aprendizaje automático también resulta beneficioso para categorizar el correo automáticamente en varias carpetas de la bandeja de entrada definidas por el usuario, como principal, social, promociones, actualización, foros [96].

Ciberseguridad: El aprendizaje automático puede ofrecer una solución a diversos problemas de ciberseguridad como; detección de ciberataques, la detección de malware, la detección de intrusiones, la seguridad de los sistemas de energía, los sistemas de control industrial, la detección de intrusiones en sistemas además de encontrar amenazas internas y mantener a salvo a las personas mientras navegan o proteger los datos en la nube al descubrir actividades sospechosas [97].

Internet de las cosas: El Internet de las cosas (IoT) convierte los objetos cotidianos en objetos inteligentes al concederles la capacidad de transmitir datos y automatizar tareas sin necesidad de intervención humana. De modo que, se considera que IoT es capaz de mejorar casi todas las actividades de nuestra vida; crea un hogar inteligente, puede intervenir en la educación, la comunicación, el transporte, el comercio minorista, la agricultura, la atención sanitaria, las empresas y muchas más. Nobakht et al. [98] mediante el uso de dispositivos inteligentes y algoritmos aprendizaje automático puede detectar actividad irregular en los hogares como la detección de intrusos, una vez que la

actividad sospechosa es detectada el sistema es capa de bloquear el acceso al hogar del intruso creando un ambiente de bajo riesgo.

Predicción del tráfico y transporte: Los sistemas de transporte son un componente fundamental del desarrollo económico de los países. Sin embargo, en varias ciudades del mundo se experimenta un aumento excesivo del volumen de tráfico, lo que desencadena: retrasos, congestión en las vías, incluso aumento del precio del combustible, además que provoca un aumento de la contaminación por CO, etc. Por lo tanto, obtener predicciones precisas del tráfico aplicando modelos de aprendizaje automático y aprendizaje profundo pueden ser de gran soporte para minimizar o contrarrestar los problemas de tráfico. Uno de los dispositivos más usados para el control de tráfico son las cámaras. Estás son tan complejas que tienen la capacidad de detectar objetos en las imágenes capturadas, de esta forma, se logra crear un ambiente seguro donde se logra detectar accidentes en las carreteras incluido los problemas en el trafico creando un ambiente inteligente y seguro en las vías [99].

Sanidad y pandemia de COVID-19: El aprendizaje automático también está presente en el campo de la medicina. Puede ser de utilidad para resolver problemas de diagnóstico y pronóstico en diversos ámbitos médicos, como ayudar en la predicción de enfermedades, la extracción de conocimientos médicos, en la detección de irregularidades en los datos, además en la gestión de pacientes, etc. El aprendizaje automático ha sido empleada en la predicción temprana de la diabetes sobre reportes médicos de pacientes (mujeres) [51]. Así mismo, mediante la aplicación de algoritmos de clasificación sobre reportes médicos digitales se consiguió soporte para la toma de decisiones médicas [100].

Comercio electrónico y recomendaciones de productos: La recomendación de productos es una forma muy conocida y aplicada del aprendizaje automático. Está presente y destacada en casi cualquier sitio web de comercio electrónico hoy en día. Aplicando aprendizaje automático se puede dar soporte a las empresas mediante el análisis de los historiales de compra de sus clientes y hacer sugerencias personalizadas de productos en su próxima compra basándose en su comportamiento y preferencias. Los algoritmos de aprendizaje automático están siendo exitosamente utilizados para las recomendaciones de libros y artículos de moda en tiendas online [101].

Reconocimiento de imágenes y patrones: El reconocimiento de imágenes es un ejemplo muy conocido de aprendizaje automático aplicado al mundo real, puede identificar un objeto a partir de una imagen digital. Tiene la capacidad de analizar una radiografía para poder etiquetarla como cancerosa o no. Incluso logra reconocer caracteres o detectar rostros en una imagen. Facebook utiliza el reconocimiento de imágenes para las sugerencias de etiquetado en las imágenes. En una investigación realizada en Indonesia se aplica el reconocimiento de imágenes sobre comidas para extraer la información de qué contienen los alimentos para usarlo como referencia sanitaria [102].

Agricultura sostenible: La agricultura es un campo esencial para la supervivencia de todas las actividades humanas. Las prácticas de una agricultura sostenible favorecen a la mejora de la productividad agrícola y a la reducción del impacto negativo que esta puede provocar en el medio ambiente. Mediante la aplicación de aprendizaje automático en la fase de preproducción, se podría predecir el rendimiento de los cultivos además de las propiedades del suelo o si existe necesidades de riego. Por otro lado, en la fase de producción se utilizaría para la predicción meteorológica o la detección de enfermedades e incluso la detección de malas hierbas o plagas. Finalmente, también serviría de soporte para la estimación de la demanda y la planificación de la producción, etc. Existe un estudio que muestra el uso de técnicas de aprendizaje automático en la agricultura para aliviar los problemas en las áreas de precosecha, cosecha y poscosecha. El aprendizaje automático enfocado a la agricultura es capa de minimizar las pérdidas, permite una agricultura más eficiente y precisa con menos mano de obra humana y una producción de alta calidad [103].

Lo escrito anteriormente evidencia la importancia de los modelos basados en el aprendizaje automático. Los problemas de aprendizaje automático cubren desde los juegos hasta los vehículos autoconducidos. El uso de ML es muy popular y está siendo ampliamente aplicado para resolver problemas de la vida real como: la bioinformática, la quimioinformática, las redes informáticas, la clasificación de secuencias de ADN, la economía y la banca, la ingeniería avanzada y muchos más [46].

2.3.4 ETAPAS DE LA MINERÍA DE TEXTO

Una gran parte de la comunicación y documentación oficial que mantienen las organizaciones comerciales, gubernamentales y otras entidades es creada en forma de documentos electrónicos textuales y correos electrónicos o incluso en la comunicación personal se utilizan correos electrónicos, blogs, etc. Es decir, el formato comúnmente usado en la vida real para almacenar información es de tipo texto, Por tanto, el texto es una fuente rica de información y debe pasar por diferentes etapas para obtener el formato computacional requerido para su estudio y así poder extraer conocimiento de él. En la Figura 21 se describe la metodología de un sistema de minería de texto organizado por sus etapas: *procesamiento, transformación, modelado y finalmente interpretación y evaluación de los resultados* [104].

Figura 21. Etapas de la Minería de Texto.

Fuente: Text mining: A Brief survey, (Patel and Soni 2012).

Como podemos observar en la Figura 21 la primera etapa es la etapa de procesamiento. Esta etapa parte inicialmente de una colección de documentos su objetivo es eliminar cualquier dato erróneo o irrelevante que no aporte valor al texto preparando a éste para obtener una mejor representación del mismo. Después del procesamiento, el texto entra en la etapa de transformación donde se normaliza el texto, se representa en un espacio vectorial, que contiene la frecuencia de las palabras en el conjunto documentos, Además en esta misma etapa se seleccionan las mejores características (palabras) que aporten valor al texto y así obtener óptimos resultados. Así mismo, en aquellos estudios donde exista un desbalance sustancial de las clases a estudiar se debe realizar un equilibrio o balanceo de las clases para evitar el sobreajuste. Seguidamente, en la fase de modelado se aplican los diferentes modelos de Aprendizaje Automático dependiendo de las características del problema a tratar: Clasificación, Regresión, Agrupación o Asociación.

Finalmente, los resultados obtenidos de la etapa de modelado se interpretan, analizan y evalúan para determinar el rendimiento de los modelos aplicados al texto [23]. A continuación, se describe de forma más detallada las etapas que envuelven la minería de texto.

2.3.4.1 PROCESAMIENTO DEL TEXTO

El procesamiento de texto incluye la aplicación de diferentes técnicas con el objetivo de limpiar (como la eliminación del ruido y de datos incoherentes o erróneos) y transformar el texto para conseguir un formato utilizable. A continuación, detallaremos algunas tareas aplicadas en el procesamiento de texto:

- **Eliminación de datos irrelevantes**: consiste en eliminar cualquier dato irrelevante como: enlace, número, signos de puntuación y caracteres especiales del texto.

- **Tokenización**: es el proceso de segmentar un texto en partes pequeñas llamadas "tokens", dicha separación se realiza mediante la detección de los espacios en blanco o los signos de puntuación.

- **Palabras vacías**: Consiste en eliminar aquellas palabras que, desde el punto de vista no lingüístico, no aportan información de importancia en los documentos. Las palabras vacías normalmente las eliminamos para ayudar a que los métodos funcionen de forma óptima, para mejorar la calidad de los datos y aumentar la eficiencia de los modelos de minería de texto.

- **Stemming**: Las diferentes formas de una misma palabra suelen ser un problema para el análisis del texto, ya que tienen una ortografía diferente y un significado similar (por ejemplo, aprende, aprendía, aprender). Para resolver este problema, se utiliza stemming, obteniendo las raíces de las palabras y sirviendo de apoyo en el proceso de reducción de dimensionalidad del texto.

2.3.4.2 TRANSFORMACIÓN DEL TEXTO

Previo a la aplicación de un algoritmo de minería de texto se debe realizar un proceso de trasformación del texto que consiste en representar las palabras en un formato óptimo, esta transformación permite consolidar los datos en la forma adecuada para su posterior estudio. En esta sección describiremos a la matriz TF-IDF, esta es una herramienta que se aplica para la transformación de texto a un formato que permita el análisis del mismo, posteriormente, se describirá a Chi-cuadrado ($\chi 2$), utilizado para seleccionar las mejores características previo al modelado, finalmente, se hablará de Smote que es una técnica usada para el balanceo de clases desequilibradas.

a) TRANSFORMACIÓN DEL TEXTO A FORMATO NÚMERICO

Para la aplicación de algoritmos de aprendizaje automático el texto debe transformarse en términos de la frecuencia relativa de presencia de las palabras en el documento y en toda la colección. Una de las técnicas utilizadas para esto es la matriz TF-IDF, esta técnica calcula una puntuación TF-IDF para cada término de un documento, basándose en su frecuencia en este documento y el número de documentos que lo incluyen [25]. Una representación de esta tarea se puede observar en la *Figura 22*. Los valores numéricos que reflejan la matriz TF-IDF representan la importancia de una palabra en un documento perteneciente a una colección de documentos. Además, TF-IDF mediante su configuración de parámetros permite filtrar y remover aquellos términos que aparecen de forma muy frecuente, asimismo, los términos que aparecen rara vez dentro del conjunto de datos, los cuales pueden ser irrelevantes en el estudio del texto.

	abandonada	acecho	agresión	casa	delito	denuncia	erradicar	familia	homicidio
0	0.9654	0.1274	0.2589	0.7896	0.4523	0.0	0.4156	0.7413	0.6541
1	0.0	0.0254	0.1478	0.9632	0.5412	0.0258	0.7532	0.3698	0.4397
2	0.4123	0.0	0.5236	0.3654	0.2587	0.6523	0.0	0.2145	0.3654
3	0.2147	0.1478	0.6541	0.0	0.7412	0.6324	0.0	0.2547	0.3214
4	0.0	0.3256	0.3579	0.1478	0.2573	0.1598	0.7452	0.0	0.6541
5	0.6213	0.4189	0.6541	0.2541	0.8523	0.0	0.3698	0.5462	0.4528

Figura 22. Representación de un espacio vectorial tfidf.

Fuente: Elaboración propia.

La representación TF-IDF es comúnmente utilizada para el tratamiento de textos, la cual consiste en la representación de las palabras en un espacio vectorial. En la representación TF-IDF la frecuencia de términos de cada palabra se normaliza mediante la frecuencia inversa del documento o IDF [27].

- TF-IDF: Aquí se transforma el texto (términos o palabras) de cada noticia en un espacio vectorial de pesos representados en formato numérico. TF-IDF combina dos métricas. TF representa el número de veces que aparece un término específico en un documento d IDF es la proporción de veces que aparece el término en todos los documentos [55]. La matriz resultante representa los términos del documento por pesos $W(d,t)$. Se computará de la siguiente manera:

$$W(d,t) = TF(d,t) \cdot \log\left(\frac{N}{df(t)}\right) \quad (6)$$

En la ecuación 1, N es el número de documentos y $df(t)$ es el número de documentos que contienen el término t. Así, el valor de TF-IDF es el resultado de la multiplicación de la frecuencia del término en un documento y el logaritmo de la cantidad de documentos que presentan ese término [105].

b) SELECCIÓN DE CARACTERÍSTICAS

El proceso de selección de características es una de las técnicas más frecuentes e importantes dentro del preprocesamiento de datos, convirtiéndose en un componente indispensable del proceso de aprendizaje automático. También se le conoce como selección de variables, selección de atributos o selección de subconjuntos de variables. Este proceso consiste en detectar características relevantes y eliminar los datos irrelevantes, redundantes o ruidosos. De esta forma, se acelera los algoritmos de minería de datos, mejora la precisión predictiva y aumenta la comprensibilidad. Se denomina características irrelevantes aquellas que no aportan información útil en el proceso de modelado [106].

Dentro de las técnicas de selección de características encontramos Chi-cuadrado ($\chi 2$). Esta técnica puede medir el grado de independencia entre un término y una categoría y se ha utilizado ampliamente para la categorización de textos [50]. Se utiliza para probar la independencia de dos eventos, una breve descripción a continuación:

- Chi-cuadrado ($\chi 2$): El texto puede contener características débiles para el modelo. Una forma de eliminarlos es el uso de la técnica Chi-cuadrado ($\chi 2$). En el proceso de selección de características, podemos encontrar dos variables. Uno es la frecuencia de ocurrencia de la característica t_i, y el otro es la probabilidad de ocurrencia de una categoría C_k. Al ejecutar el método Chi-cuadrado ($\chi 2$) , podemos examinar la relevancia entre *ti* y la categoría C_k. Cuanta más información relevante tenga *ti*, mayor puntuación tendrá C_k. Se puede definir de la siguiente manera:

$$\mathrm{X}^2(t_i, C_k) = \frac{N(ad - cb)^2}{(a + c)(b + d)(a + b)(c + d)} \quad (7)$$

En la ecuación a representa el número de veces que ocurren t y C. Por otro lado b es el número de veces que aparece t sin C. c es el número de veces que aparece C sin t. Asimismo d es el número de veces que no aparece C y tampoco t. Finalmente, N es el número total de documentos [107][55]. El resultado de esta ecuación es la correlación

entre la característica t_i y la categoría C_k . Si es 0, entonces son independientes. Cuanto mayor sea el valor, más relación tiene t_i con la categoría [108].

c) EQUILIBRADO DE DATOS

Los modelos de aprendizaje automático tienen la capacidad para predecir o clasificar objetos y datos mediante el aprendizaje de patrones del conjunto de datos. El rendimiento del clasificador no sólo depende del modelo de aprendizaje del algoritmo, sino que la composición de cada clase y la distribución de los datos desempeñan un papel importante [109]. Sin embargo, en la vida real la existencia de conjuntos de datos con clases desequilibradas es bastante común. En este tipo de datos existe una disparidad sustancial en las distribuciones de clases, lo que supone retos notables en el campo del aprendizaje automático. Algunos ejemplos de estos casos son: los datos de los ámbitos médico o financiero suelen estar desequilibrados porque ciertos tipos de instancias (también llamados ejemplos o muestras) son naturalmente raros, como los pacientes de cáncer o las transacciones anómalas.

Un conjunto de datos puede ser considerado desequilibrado si al menos una de las clases posee una cantidad pequeña de instancias en comparación con las demás clases [110]. Cabe mencionar que, un conjunto de datos puede estar desequilibrado debido a diversos factores, como la rareza de determinados sucesos o la distribución desigual de los casos en la población real. Por lo tanto, es crucial comprender las implicaciones de los conjuntos de datos desequilibrados y diseñar estrategias para tratar este problema con eficacia y así obtener un buen desempeño de los algoritmos.

SMOTE es uno de los métodos pioneros para abordar el desequilibrio de clases. Es un método de sobremuestreo que se utiliza para remediar los problemas causados por la clasificación de conjuntos con datos desequilibrados [111]. El algoritmo SMOTE puede mejorar el efecto de clasificación de los datos desequilibrados generando aleatoriamente nuevos puntos de muestra minoritarios para aumentar hasta cierto punto el índice de desequilibrio [112]. Para reequilibrar el conjunto de entrenamiento original SMOTE crea nuevos datos minoritarios interpolando varias instancias minoritarias, las nuevas instancias creadas se denominan ejemplos "sintéticos". SMOTE, sobremuestrea buscando

puntos de datos de la misma clase que una muestra para hacer una línea de interpolación y formar nuevos datos a lo largo de la línea. Se generan muestras de datos sintéticos junto con los segmentos de línea conectados a algunos o todos los vecinos k más cercanos de la clase minoritaria. Para encontrar los puntos de muestra vecinos más cercanos utiliza el algoritmo K vecinos más cercanos (KNN) mediante el cálculo del valor de la distancia entre los vectores de características (muestras) y sus vecinos más cercanos en la clase minoritaria. Los vecinos del vecino k más cercano luego se multiplican con un número aleatorio entre 0 y 1 y se añaden al vector de características previamente seleccionado. La distancia de los puntos de datos de la muestra a otros puntos de datos se miden utilizando el método de medición de la distancia euclídea [109].

El método Smote ha sido aplicado demostrando una significativa mejora del rendimiento de los clasificadores. Chawla et al. [113] investigó el impacto de SMOTE en varios algoritmos de clasificación y observó sistemáticamente mejoras en la precisión de la clasificación y en la medida F1 (La puntuación F1 mide la media armónica de la precisión y la recalificación, lo que sirve para medir la eficacia derivada) para la clase minoritaria. En otra investigación se examinó el rendimiento de SMOTE aplicándolo en diversos conjuntos de datos desequilibrados y sus autores llegaron a la conclusión de que mejora el rendimiento general del clasificador, en particular en escenarios con desequilibrio grave de clases. Seiffert et al. [114] los autores realizaron una evaluación exhaustiva de los métodos de sobremuestreo, incluido SMOTE, e informaron sistemáticamente de su superioridad a la hora de mejorar la precisión de la clasificación para la clase minoritaria y el rendimiento predictivo general. Hamid et al. [115] señaló que SMOTE contribuye a determinar mejor los límites de decisión y mejora la capacidad de generalización de los modelos de clasificación. A partir de lo anterior, SMOTE demuestra su adaptabilidad y eficacia a través de una amplia gama de aplicaciones.

Se debe estacar, la aplicación de un método para equilibrar las clases del conjunto de datos es imprescindible, mediante su uso se minimiza el coste global de la clasificación errónea, por lo que, SMOTE es una valiosa herramienta en una valiosa herramienta para mitigar los problemas de desequilibrio, contribuye a mejorar la precisión de los modelos y a obtener predicciones fiables en diversos ámbitos.

2.3.4.3 MODELADO Y EVALUACION

Una vez el texto procesado y transformado para su utilización, nos encontramos con las etapas de modelado y evaluación. La etapa de modelado consiste en aplicar una serie de algoritmos para realizar una determinada tarea orientada a conseguir un determinado objetivo. En concreto en nuestro caso, trataremos la clasificación de textos, extracción de temas latentes y reglas de asociación. Por otro lado, la etapa de evaluación consiste en analizar y evaluar los resultados obtenidos de la etapa de modelado, para así, poder determinar el rendimiento del algoritmo aplicado.

A. CLASIFICACIÓN DE TEXTO

Las bases de datos son ricas en información oculta que pueden utilizarse para hacer un análisis o estudios para así ayudar en la toma de decisiones inteligentes. Con los algoritmos de categorización o clasificación se puede asignar automáticamente a un documento de texto un conjunto de clases predefinidas. La categorización suele basarse en un tesauro, los temas predefinidos, y las relaciones se identifican mediante la búsqueda de términos generales, más específicos, sinónimos y términos relacionados. Las herramientas de categorización suelen disponer de un método para clasificar los documentos en función de su mayor contenido sobre un tema concreto [116]. La clasificación de textos tiene varios campos de aplicación [29], podemos citar los siguientes:

- ***Clasificación de artículos de noticias:*** Este es un proceso en el que se clasifica una amplia cantidad de documentos como investigaciones científicas, registros legales, artículos de noticias. Por ejemplo, un artículo de noticia puede ser clasificado dentro de las clases: deportes, finanzas, seguridad, salud, tecnología, etc.

- ***Filtrado automático de correos:*** Una de las utilidades de la minería de texto es que con la aplicación de las técnicas de clasificación se puede exitosamente detectar y filtrar correos no deseados.

- ***Clasificación de páginas web:*** En esta tarea se clasifican páginas web dentro de predefinidas categorías basadas en su contenido.

- ***Desambiguación de sentido de las palabras:*** En esta tarea se identifican una serie de fenómenos lingüísticos como "selección preferencial o información de dominio" que es relevante para resolver la ambigüedad de las palabras.

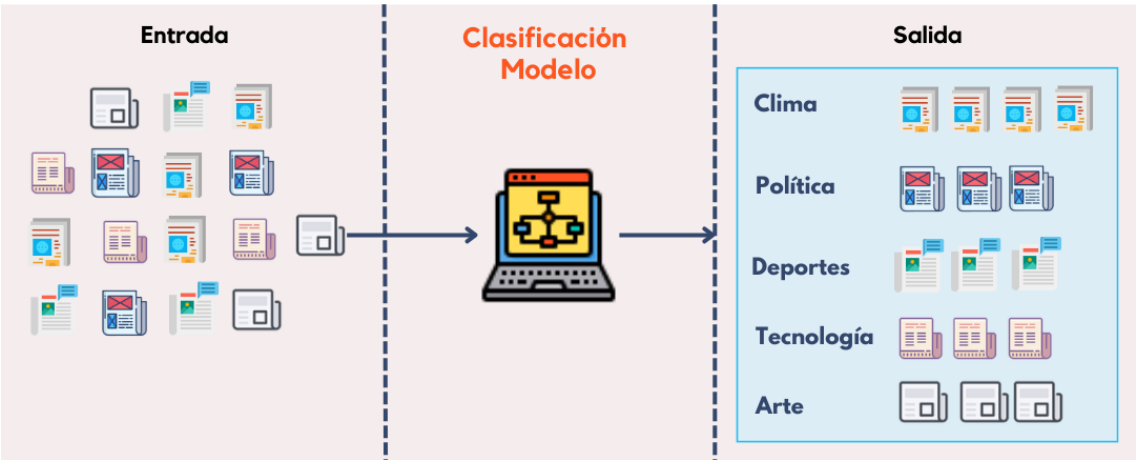

Figura 23. Representación de la clasificación de texto.

Fuente: Elaboración propia.

En la *Figura 23* se describe una representación de un proceso de clasificación de texto. A partir de una colección de texto y mediante la aplicación de un modelo de clasificación se logra clasificar cada documento dentro de 5 categorías previamente definidas: (Clima, Política, Deportes, Tecnología, Arte). Cada etiqueta asignada está relacionada con el contenido del texto.

A.1 MEDIDAS DE EVALUACIÓN DE UN PROCESO DE CLASIFICACIÓN

Existen varias formas de medir la eficacia de un clasificador, entre las medidas más utilizadas tenemos la "Accuracy". Otras medidas alternativas aplicadas en la clasificación de texto son; "*Precisión y F1 score"*. la puntuación *F1* y el *Accuracy*. Estas métricas se derivan de determinar si un documento fue verdadero positivo (TF), falso positivo (FP), verdadero negativo (TN) o falso negativo (FN). En nuestra investigación adoptamos las medidas estadísticas antes mencionadas para evaluar el desempeño de la clasificación realizada por los modelos, al igual que, en los trabajo: [117]–[119].

La "*Accuracy*" se puede traducir como "precisión" es un indicador que muy a menudo se aplica para evaluar el desempeño de los clasificadores que trabajan en problemas basados en texto [108]. Esta métrica representa la proporción de documentos clasificados correctamente dentro del conjunto de datos. Es una medida más general porque calcula el número de clasificaciones correctas en el conjunto de datos. Ha sido ampliamente utilizada [121]–[123] y se puede calcular mediante la siguiente fórmula [108][55][123]:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (8)$$

Por otro lado, "*F1 score*" es una métrica alternativa que equilibra la *"precisión"* (también conocida como especificidad o tasa de verdaderos negativos) y la *"recall"* (también conocida como sensibilidad o tasa de verdaderos positivos) [124]. La "*F1 score*" representa la media armónica de *"precision* y *recall"* (recall es la proporción de casos positivos detectados correctamente por el clasificador). Como resultado, de esta métrica el clasificador sólo obtendrá una puntuación "*F1 score*" alta si tanto la precisión como la recuperación son altas. Esta medida representa la capacidad de los clasificadores para ubicar un documento dentro de la clase correcta [55]. Esta métrica se puede definir de la siguiente manera:

$$F1\ score = \frac{2TP}{2TP + FP + FN} \quad (9)$$

Por último, la "*precisión*" es una métrica que mide el número de muestras correctamente clasificadas frente al total de muestras. La "*precisión*" se calcula mediante la relación entre el total de clasificaciones correctas y el total de clasificaciones realizadas [55]. Una precisión igual a 1 indica un porcentaje del 100% del clasificador y que todas las instancias se clasificaron correctamente [55]. Esta medida se obtiene con la sigue fórmula:

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Cabe mencionar que *"Accuracy"* también puede definirse como el grado de aproximación a un determinado valor esperado. Se refiere a la calidad del resultado y no es equivalente

a la precisión [125]. Por lo que, una estimación cuantitativa "*Accuracy"* es esencial para definir el grado de confianza que puede depositarse del modelo y la fiabilidad de las decisiones basadas en dicho resultado [126]. Mientras que la *"precisión"* es el término cualitativo que designa la concordancia entre resultados de pruebas independientes obtenidos en condiciones estipuladas[125]. Un gran número de autores han aplicado las métricas antes mencionadas para evaluar los resultados obtenidos con una amplia variedad de algoritmos de aprendizaje automático; [43], [51], [121], [122], [127]–[133].

B. MODELADO DE TEMAS LATENTES

El modelado de temas es un método probabilístico que a partir de un análisis jerárquico bayesiano de los textos originales puede descubrir la estructura semántica subyacente de una colección de documentos. Ha sido aplicado a muchos tipos de documentos: correos, documentos científicos, incluso noticias. Siendo capaz de descubrir patrones de palabras y conectar documentos que presentan patrones similares, el modelado de temas ha surgido como una nueva y poderosa técnica para encontrar una estructura útil en una colección de documentos que no tiene un formato estructurado [134].

Otro rasgo del modelado de temas es que está dentro del campo del procesamiento de lenguaje natural, es un modelo estadístico que trata de determinar o reconocer los patrones o ámbitos en común de las palabras que aparecen en un texto. Lomoschitz [135] en su investigación expone que en el proceso de modelado de temas inicialmente se debe proporcionar un "corpus" que es una colección de documentos con alguna relación entre sí, como ejemplo, poseer una estructura similar, referirse sobre el mismo. En la etapa del modelado el sistema procesa los documentos, analiza y determina en qué documentos suelen aparecer las palabras, y en cuáles suelen reaparecer las mismas. Los grupos de palabras que hagan referencia a documentos similares conformarán un "tema". Los temas generados se definen como una mezcla probabilística de palabras: unos términos aparecerán con muy altas probabilidades dentro de un tema, mientras que otros tendrán una probabilidad muy baja o casi nula por no ser característicos para ese tema [119].

Las técnicas de Modelado de Tema pertenecen a los modelos no supervisados, puesto que no hay un conjunto de datos pre-etiquetado para entrenar al modelo. El modelado de

temas genera un resultado, el experto lo evalúa y define si los resultados son óptimos, si se mantiene el modelo o si se pueden optimizar los resultados implementando otros parámetros que mejoren el rendimiento del modelo[119].

Previo, al proceso de Modelado se debe seleccionar el número de temas que se desea obtener. El modelo generado revelará los patrones encontrados en los documentos del corpus, y proporcionará una lista de los topics encontrados juntos con sus términos más significantes, así como la distribución de topics de los documentos (probabilidades de cada temática de aparecer dentro de cada documento) [119].

Por otro lado, Maryamah et al. [136] explica que el modelado de temas es un método de búsqueda de temas en datos no estructurados con el proposito de determinar la palabra dominante que será el tema de los datos. El modelado de temas puede realizarse aplicando métodos como Latent Dirichlet Allocation (LDA). Cabe mencionar que, LDA es tiene una alta capacidad de determinar patrones o tendencias en cualquier campo. Sharma et al. [84] menciona que LDA ha sido aplicado para el análisis de los datos de temas candentes o tendencias publicadas en Twitter, el análisis de reseñas de Internet, la agricultura, la ingeniería de software, el medio ambiente, el aprendizaje profundo, la medicina y muchos más. Así mismo, existen estudios donde se ha demostrado la utilidad de LDA analizando texto corto [83], [137], [138]. Por otro lado, también se puede encontrar poca evidencia sobre la aplicabilidad y los buenos resultados que genera el modelado de temas trabajando sobre texto extenso [119], [139], [140].

C. EXTRACCIÓN DE REGLAS DE ASOCIACIÓN

A través de la aplicación de extracción de reglas de asociación podemos encontrar relaciones entre los elementos que contiene un texto. Es una de las tecnologías más utilizadas en el campo de la minería de datos. Fue introducida por Agrawal et al. [141] y es utilizada habitualmente para determinar patrones ocultos y relaciones entre las variables de los conjuntos de datos y las dependencias entre estas variables. Los algoritmos de reglas de asociación se han aplicado con éxito en diversos campos, como la industria sanitaria, el análisis de cestas de mercado y los sistemas de recomendación [85].

El objetivo de los modelos de reglas de asociación es extraer un conjunto de características fuertemente correlacionadas que comparten un gran número de registros dentro de una base de datos determinada. Por ejemplo, las reglas de asociación halladas en una base de datos de ventas pueden ser útiles para la toma de decisiones de los responsables de marketing. Se han mostrado también bastantes útiles también en el análisis de la cesta de la compra , descubriendo asociaciones entre varios artículos y obteniendo resultados que han ayudado a los responsables de la toma de decisiones a analizar los hábitos de compra de los clientes [142].

Las reglas de asociación son un método de aprendizaje no supervisado muy significativo para el reconocimiento de patrones. Las reglas de asociación no exigen que los datos de la muestra obedezcan a una distribución normal, cumplan la prueba de correlación o sean continuos, por lo que son más adaptables [143]. Su aplicación permite descubrir asociaciones en datos llamados "transacciones", de esta forma, podemos identificar relaciones entre palabras y analizar su comportamiento.

Las reglas de asociación se utilizan para obtener conocimientos relevantes a partir de grandes bases de datos transaccionales. Una base de datos transaccional podría ser, por ejemplo, una base de datos de cestas de la compra, donde los artículos serían los productos, o una base de datos de texto donde los artículos son las palabras. Una regla de asociación se representa de la forma X→Y donde X es un elemento que representa el antecedente e Y un elemento que representa el consecuente. Como resultado, se concluye que los elementos consecuentes tienen una relación de coocurrencia con los elementos antecedentes. Por lo tanto, las reglas de asociación pueden utilizarse como método de extraer relaciones ocultas entre elementos o ítems dentro de bases de datos transaccionales.

En la *Figura 24* se representa un ejemplo de reglas de asociación extraídas de los productos que compran los clientes en el supermercado. A partir del análisis de los productos dentro de las transacciones se determinó las reglas: *"SI Pan ENTONCES Leche"* además *"SI Flores ENTONCES Vino"*, Las reglas que se obtienen tienen una forma de si-entonces, lo que indica que existe una relación entre esos artículos, si se compra *Pan* también se compra *Leche*, y si se compra *Flores* también se compra *Vino*.

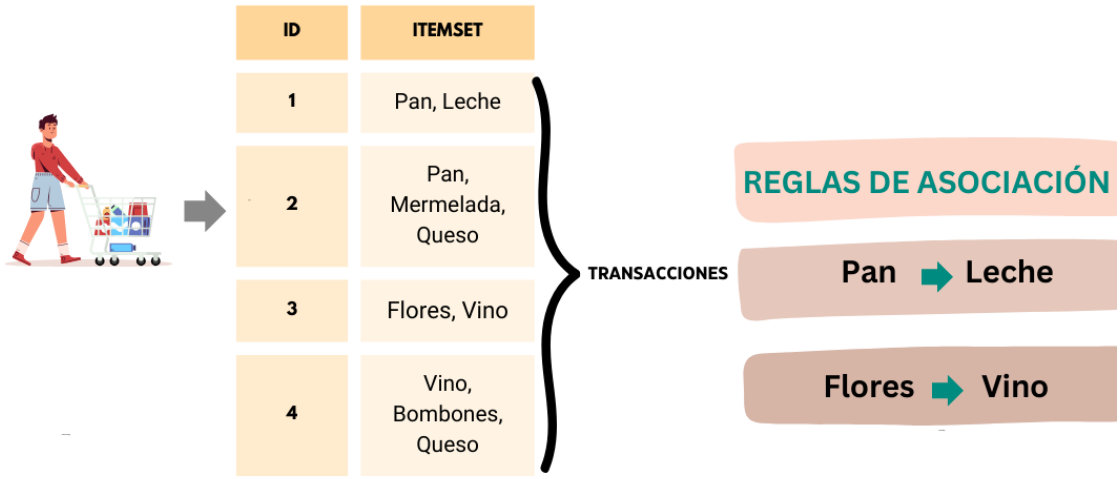

Figura 24. Ejemplo reglas de asociación.
Fuente: Elaboración propia.

Si nos centramos en nuestro tema, las reglas de asociación se pueden aplicar sobre texto consiguiendo extraer reglas con el formato *SI* ⇒ *ENTONCES*. Una asociación habitual es la asociación de palabras, aquí las reglas de asociación se generan a partir de conjuntos de palabras. Por otro lado, la asociación de textos, denominada asociación de documentos o artículos es aquella en la que las reglas de asociación de artículos o documentos se generan a partir de conjuntos de documentos. Por último, la asociación de frases o párrafos, denominada asociación de subtextos, es aquella en la que las reglas de asociación de frases o párrafos se generan a partir de conjuntos [144].

C.1. MEDIDAS DE EVALUACIÓN DE LOS PATRONES DE ASOCIACIÓN GENERADOS

En la minería de reglas de asociación la forma clásica de medir la bondad de las reglas de asociación en relación con un problema dado es con tres medidas: soporte, confianza y elevación, que se definen como sigue [145]:

- **Soporte de un conjunto de elementos:** Se representa como supp (X), y es la proporción de transacciones que contienen el elemento X sobre la cantidad total de transacciones del conjunto de datos (D). La ecuación para definir el soporte de un conjunto de artículos es:

$$supp\ (X) = \frac{|t \in D : X \subseteq t|}{|D|} \quad (11)$$

- **Soporte de una regla de asociación:** Se representa como supp (X → Y), es la cantidad total de transacciones que contienen ambos elementos X e Y, como se define en la siguiente ecuación:

$$supp\ (X \rightarrow Y) = supp\ (X \cup Y)\ \ (12)$$

- **Confianza de una regla de asociación:** Se representa como conf (X→Y) y representa la proporción de transacciones que contienen el elemento X que también contiene Y. La ecuación es:

$$conf\ (X) = \frac{supp\ (X \cup Y)}{supp\ (X)}\ \ (13)$$

- **Elevación:** Es una medida útil para evaluar la independencia entre los elementos de una determinada regla de asociación. La medida lift (X→Y) representa el grado en que X es frecuente cuando Y está presente o viceversa. La elevación se define matemáticamente de la siguiente manera:

$$lift\ (X\ \rightarrow Y) = \frac{conf\ (X \rightarrow Y)}{supp\ (Y)}\ (14)$$

3. METODOLOGÍA

En este capítulo nos centraremos en describir los pasos que hemos llevado a cabo en nuestra investigación sobre el estudio de la violencia contra la mujer. Como hemos comentado en secciones anteriores nuestro problema consiste en un problema de minería de textos. Por lo que, se desarrolló un proceso de clasificación de noticias en tipos violencia, de igual forma, un proceso para la generación de temas latentes, por último, un proceso para la extracción de patrones sobre la VCM. A continuación, detallaremos las tareas realizadas en cada uno de los procesos que han sido citados.

3.1 CLASIFICACIÓN DE NOTICIAS EN TIPOS DE VCM

Como es conocido, la clasificación de texto es una técnica del procesamiento de lenguaje natural que aplica técnicas de aprendizaje automático para identificar a qué categoría o clase pertenece un texto. En esta sección se detallarán las actividades realizadas en el proceso de clasificación de noticias en diferentes tipos de violencia contra la mujer. Inicialmente, se citará cómo se realizó la recolección del texto, seguido por el preprocesamiento del texto, la transformación del texto. Finalmente, hablaremos de los modelos aplicados y la evaluación de los resultados obtenidos, en la *Figura 25* se puede observar las etapas del proceso de clasificación junto a las tareas realizadas en cada etapa de la clasificación de noticas [119].

Figura 25. Metodología aplicada para la clasificación de texto.

Fuente: Elaboración propia.

3.1.1 RECUPERACIÓN Y PROCESAMIENTO DEL TEXTO

En esta sección inicialmente se detallarán las actividades realizadas para la recuperación y colección del conjunto de noticias sobre la violencia contra la mujer que se va a estudiar, seguido por las tareas de procesamiento del texto, las cuales limpiaran y prepararan el texto para obtener una mejor representación del mismo.

A. RECUPERACIÓN DEL TEXTO

Para el desarrollo de esta investigación se empezó con la recuperación de noticias sobre la Violencia Contra la Mujer de periódicos digitales mediante un proceso de raspado web o "web scraping". Las técnicas de raspado web permiten recoger información de diferentes fuentes: páginas web, emails, documentos de texto, PDF, reportes, audios y videos que hayan sido publicados en internet. Nosotros aplicamos "raspado web" para la recopilación de noticias sobre la VCM publicados en periódicos digitales bajo el formato HTML. Se analizó la estructura HTML de la página web que contenía la noticia para identificar el o los nodos HTML que contienen la información requerida. Fue preciso

analizar la estructura de los sitios web oficiales de cada diario ya que algunos poseían diferente estructura HTML y se corría el riesgo de descargar la información incorrecta. El proceso de recolección consistió en enviar una petición a la página web de la noticia; si el destinatario procesa y acepta la petición los datos serán obtenidos de su fuente y serán enviados al programa de raspado que lo solicitó, para su posterior almacenamiento [20], [119].

Este estudio fue aplicado a textos escritos en idioma español. Debido a esto el proceso de raspado web se realizó en los periódicos digitales: Perú21, El Salvador, Tiempo, Unomasuno, 24horas, Al día, BBC Mundo, El Comercio, Extra, Milenio, Mundo y Nación, los cuales poseen su redacción en el idioma requerido. Estas fuentes de información pueden contener noticias de diferentes índoles, por lo que, para extraer noticias específicas sobre la VCM nosotros definimos varias palabras claves que reflejen un acto de violencia contra la mujer, de esta forma, se haría una búsqueda filtrada de noticias referentes a la VCM, entre esos términos tenemos *femicidio, feminicidio, violencia de género, mujer asesinada, violencia doméstica*. De esta forma en el año 2020 coleccionamos 7000 noticias (periódicos digitales) sobre la VCM en un formato no estructurado, posteriormente se procedió a almacenar la información como una colección de documentos en MongoDB, la cual es una base de datos no relacional especializada para contener grandes volúmenes de datos almacenándolos como colecciones [119].

Para el desarrollo del script encargado de la recuperación de las noticias desde internet se aplicó la librería Rvest que contiene Rstudio. Aquí se envían los parámetros de entrada (url, palabras claves, y secciones html que contienen el texto deseado), y se empieza con la solicitud de búsqueda en una fuente web (periódicos digitales) de todos los recursos (noticias) que coincidan con las palabras clave. Al coincidir, se recoge el contenido de los nodos para que podamos almacenarlo en un marco de datos en este caso, como ya hemos comentado utilizaremos MongoDB [119].

En la *Figura 26* podemos observar la distribución de las noticias descargadas, el periódico digital con el que se obtuvo una mayor concentración de datos fue el diario la nación, por el contrario, el diario con menos concentración de noticias fue el diario El Salvador [119].

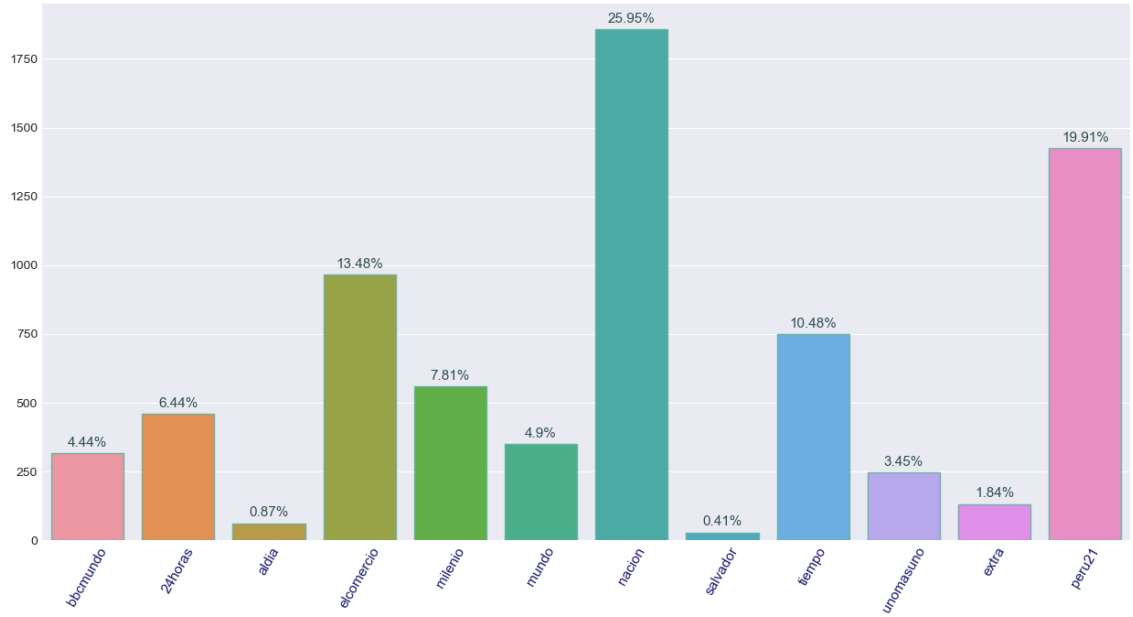

Figura 26. Distribución de noticias por fuente.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la *Figura 27* se puede observar que nuestro proceso de raspado web logró recoger documentos que fueron publicados entre los años 2002 hasta el 2020. Obteniendo la mayor concentración de datos entre los años 2016 y 2019 [119].

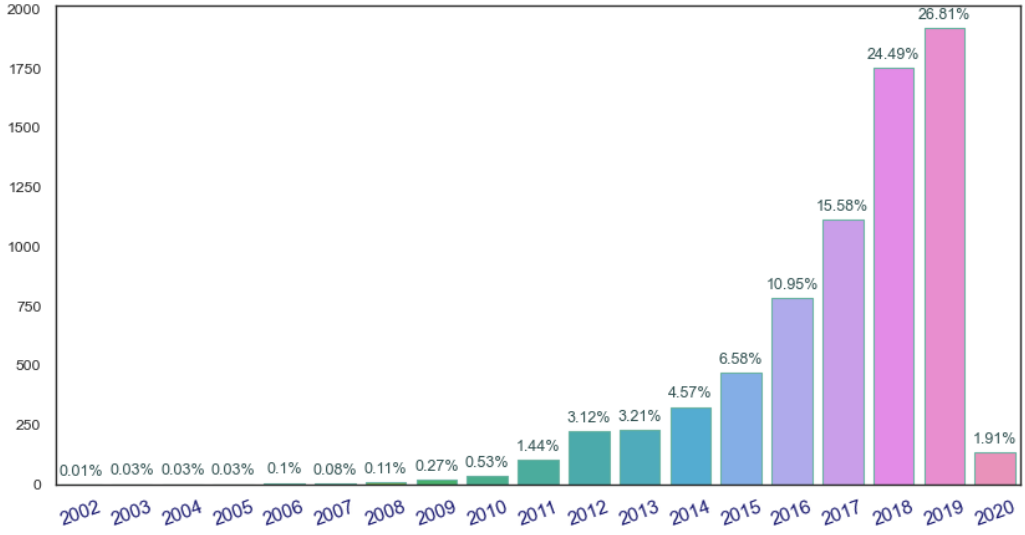

Figura 27. Distribución de datos por año.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

B. PROCESAMIENTO DEL TEXTO

Cómo se conoce, el tratamiento de los datos juega un rol muy importante para la aplicación de técnicas de minería de texto [146]. Esto se convierte en una tarea imprescindible para el descubrimiento de conocimiento cuyo objetivo es dar estructura a los datos sin formato definido. En este estudio el tratamiento de texto consistió de las siguientes tareas [65]:

- **Limpieza del texto**: Con el proceso de limpieza eliminamos términos innecesarios dentro de los documentos que pueden afectar el rendimiento de los modelos de minería de texto. Inicialmente, en esta tarea se realizó una limpieza eliminando cualquier enlace, número, y caracteres especiales del texto, en algunos documentos se pudieron encontrar emoticones, también fueron eliminados. Además, así reducimos el tamaño de los documentos. En la *Figura 28* podemos observar que un documento que contenía una dimensión de 1800 palabras (tokens) posterior al proceso de limpieza su dimensión cambia debido a que se han eliminado palabras no relevantes [119].

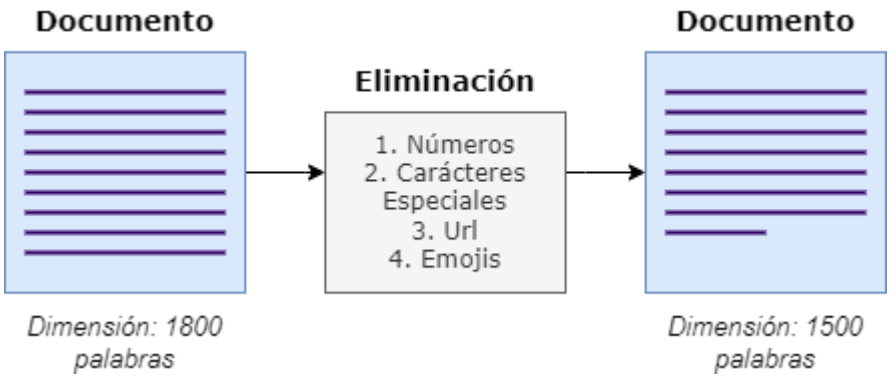

Figura 28. Limpieza del texto.
Fuente: Elaboración propia.

- **Palabras Vacías:** Consiste en la eliminación de palabras que no aportan importancia en los documentos. Las palabras vacías son palabras gramaticales irrelevantes para el contenido del texto, por lo que, se deben eliminar para mejorar la calidad de los datos y aumentar la eficiencia de los modelos de minería de texto. En esta tarea se aplicó la librería NLTK, la misma que viene con una lista por defecto de palabras vacías para el idioma español. Sin embargo, NLTK no logró

eliminar por completo las palabras vacías, con lo que se añadieron más palabras a la lista de palabras vacías por defecto. Estas palabras consistían en: términos muy frecuentes dentro de la colección de documentos como: "femicidio y feminicidio", las cuales aparecían de forma muy recurrente en el texto". De igual forma, otras palabras de poca relevancia "hubo, hizo, mujer, había, entre otros". Estos términos fueron eliminados debido a que no añaden valor al texto y pueden afectar el proceso de aprendizaje de los modelos de minería de texto. En la *Figura 29* podemos observar el proceso de eliminación de las palabras vacías al igual que el proceso de limpieza al eliminar las palabras vacías la dimensión del documento se reduce, podemos observar que un documento con una dimensión de 1500 palabras pasa a contener 1200 palabras [119].

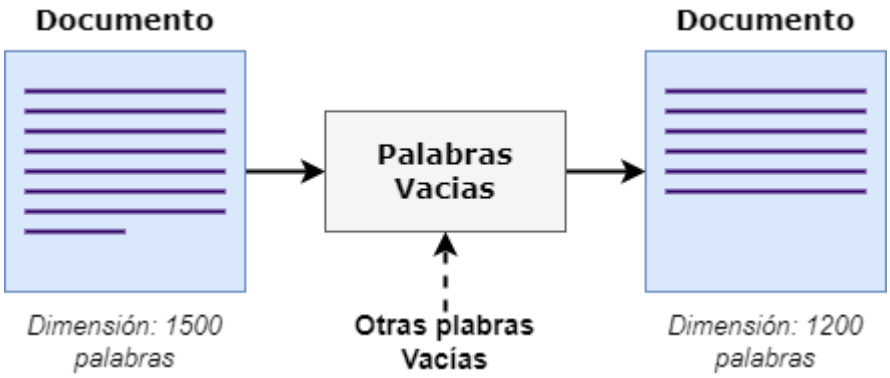

Figura 29. Proceso de palabras vacías.
Fuente: Elaboración propia.

• **Tokenización:** Esta tarea se define como el proceso de segmentar un texto en tokens mediante los espacios en blanco o los signos de puntuación. Aquí se utilizaron los documentos (noticas) como datos de entrada, posteriormente se genera la lista de tokens como salida del proceso. En la *Figura 30* se ilustra el proceso y la transformación que sufre un documento. En este proceso la cadena de palabras que contiene el documento serán representadas por palabras sueltas [119].

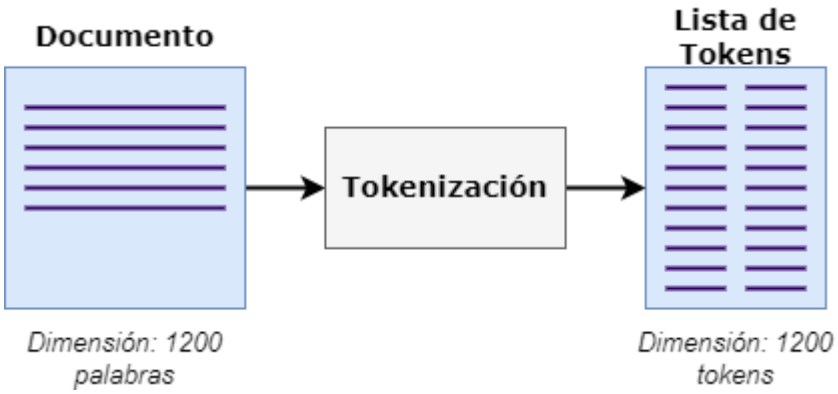

Figura 30. Proceso de Tokenización.
Fuente: Elaboración propia.

- **Stemming:** Esto refiere al proceso de mapear cada token en su propia raíz. En esta tarea reemplazamos cada término por su raíz, muchas de las palabras existentes en un texto se pueden derivar de la misma raíz gramatical. Al aplicar stemming se contribuye a la reducción de la dimensionalidad de los documentos. En la *Figura 31* podemos observar cómo funciona este proceso, dado una lista de tokens (palabas) estas sufren transformaciones donde se obtiene la raíz de cada token.

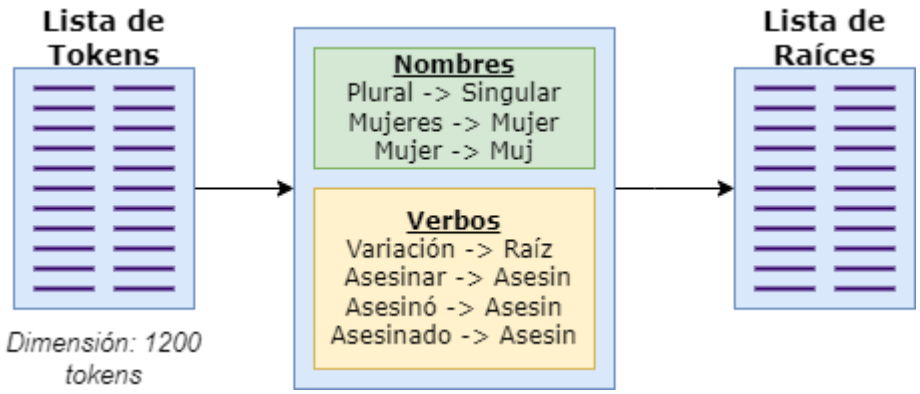

Figura 31. Proceso para obtener las raíces de los tokens.
Fuente: Elaboración propia.

3.1.2 TRANSFORMACIÓN DEL TEXTO

Una vez recuperado la colección de 7000 documentos y habiendo realizado el preprocesamiento del texto, se deben realizar una serie de procesos que permitan la aplicación de algoritmos de aprendizaje automático. En esta sección se describirán las tareas aplicadas para preparar el texto y obtener una mejor representación que facilite su

análisis computacional. Inicialmente se etiquetó el conjunto de entrenamiento, luego se transformó el texto en un espacio vectorial que represente las palabras de acuerdo a su frecuencia en el texto. A continuación, se realizó la selección de características y finalmente se equilibraron las clases a predecir [119].

A. ETIQUETADO DEL CONJUNTO DE ENTRENAMIENTO

La clasificación de texto consiste en clasificar o agrupar los documentos de texto en una o varias etiquetas de clases predefinidas, estas clases deben ser capaces de describir la idea principal del documento basado en su contenido. En esta investigación los documentos fueron clasificados en los diferentes tipos de violencia que puede sufrir la mujer. Previo al proceso de aprendizaje automático para la clasificación de texto es necesario poseer un conjunto de datos pre-etiquetado con las clases a predecir también llamado partición de entrenamiento, en este caso, nuestro conjunto de entrenamiento representó el 70% del conjunto de documentos. Dicho subconjunto de documentos se empleó en el entrenamiento para los modelos de clasificación. Igualmente, en el proceso de modelado es imprescindible poseer un subconjunto de validación para verificar la bondad del ajuste, en este caso, el subconjunto de validación se conformó del 30% del conjunto de documentos. En el subconjunto de validación se aplicaron los conocimientos adquiridos por el modelo y asignó una clase (tipo de violencia) a los documentos nuevos. En esta sección, describimos el proceso de creación de diccionarios sobre los tipos de violencia y los métodos de etiquetado aplicados al conjunto de entrenamiento, previo a la clasificación de texto [119].

i. CREACION DE DICCIONARIOS Y ASIGNACION DE PESOS

Para el proceso de etiquetado del conjunto de entrenamiento se crean diccionarios para cada tipo de violencia, estos diccionarios contienen términos que definen cada tipo de violencia contra la mujer (VCM), además, estos términos tendrán un peso asignado, dicho peso reflejará la magnitud e impacto de dicho término. Asimismo, habrá términos de los diccionarios que serán considerados cómo términos fuertes. Los términos fuertes son aquellos que expresan una violencia de mayor impacto. En la *Figura 32* se observa cómo se organizó y relacionó los tipos de violencia, diccionarios, pesos y palabras fuertes [119].

Si observamos la *Figura 32* podemos ver que describe la estructura de los diccionarios. Los tipos de violencia a estudiar son: *Física, Sexual y Psicológica*. Por consiguiente, se crearon diccionarios para estos tipos de violencia asignándole a cada uno de ellos los términos que describen el tipo de violencia correspondiente. También, se le asignó a cada termino un peso y se detectaron los términos fuertes que son aquellos que expresan un acto de violencia de alto impacto [119].

Diccionarios Tipos de violencia	Términos del diccionario	Pesos de los términos	Palabras fuertes del diccionario
• FÍSICA	• Decapitar • Ahorcar • Patear • -- • -- • --	• 0.08 • 0.08 • 0.03 • -- • -- • --	• Decapitar • Ahorcar • -- • -- • --
• PSICOLÓGICA	• Abuso verbal • Discriminación • -- • -- • --	• 0.06 • 0.07 • --- • --- • ---	• Discriminación • -- • -- • --
• SEXUAL	• Violación • Incesto • -- • -- • --	• 0.07 • 0.08 • --- • --- • ---	• "en este diccionarios todos los términos fueron considerados como fuertes"

Figura 32. Organización diccionarios, pesos y términos fuertes.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

Los términos de los diccionarios fueron seleccionados a partir de un tesauro sobre la VCM y de un análisis exploratorio de la frecuencia de palabras realizada a la colección de documentos. Para asignar peso a los términos se usó la teoría probabilística de ocurrencia donde: "*La probabilidad de que ocurra un evento específico oscila entre 0 y 1, donde 0 significa que no sucederá y 1 que sí sucederá*". Dicho lo anterior, el valor probabilístico de ocurrencia de un determinado tipo violencia en un documento tendrá un valor de 1. En consecuencia, el valor 1 se asignará entre cada uno los términos que contenga el diccionario. Cuanto mayor sea el impacto que representa el término del diccionario, mayor peso se debe asignar. La sumatoria de los pesos de los términos de un diccionario especifico tiene que ser igual a 1 (valor que representa la existencia de un tipo de violencia). En este caso existen 3 diccionarios; violencia física, psicológica y sexual; por

lo que el proceso de distribución del valor de ocurrencia 1 será realizado a cada diccionario y cada uno de sus términos [119].

Además, del conjunto de términos que conforman los diccionarios se identificó un subconjunto de términos que serán denominados como términos fuertes para dicho diccionario. En el caso de violencia física (F), se asignaron como términos fuertes aquellos que describen la muerte de la víctima, ejemplo: degollar, decapitar, etc. En el caso de violencia psicológica (P), las palabras racistas y discriminatorias fueron asignadas como fuertes, ejemplo: discriminación. A excepción, en la violencia sexual no se creó un subconjunto de términos fuertes debido a que todos los términos representan una coacción carnal, por lo que, todos fueron consideradas fuertes [119].

Tabla 1.Distribución de los diccionarios.

	Diccionarios	
Tipo de Violencia	*Número total de términos*	*Número de términos fuertes*
Física	1.....43	1.....7
Psicológica	1.....25	1.....8
Sexual	1.....8	1.....8

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la Tabla 1 se detalla la estructura de los diccionarios construidos. El diccionario Violencia Física tiene 43 términos de los cuales 7 términos son asignados como términos fuertes. El diccionario el psicológico tiene un total de 25 términos de los cuales 8 fueron fuertes. En el caso del diccionario sexual este presenta 8 términos, pero como son de naturaleza sexual, todos se fuertes [119].

ii. ETIQUETADO BASADO EN COINCIDENCIAS ENTRE LOS TÉRMINOS DE LOS DICCIONARIOS Y LOS TÉRMINOS DE LOS DOCUMENTOS.

Una vez que creamos los diccionarios con un conjunto de términos, con sus respectivos pesos para los términos y sus términos relevantes, pasaremos a etiquetar cada uno de los documentos. Este proceso de etiquetado consistió en calcular las coincidencias existentes

entre los términos de los diccionarios y los términos de los documentos. Cabe mencionar que, un documento puede tener coincidencias con uno o más diccionarios pudiendo ser etiquetado con una o varios tipos de violencia. Como consecuencia, en nuestra colección de 7000 documentos (noticias) para asignar una clase (tipo de violencia) a un documento se obtuvo la opinión de un experto y se determinó: *que se debe cumplir al menos una de las siguientes reglas para etiquetar un documento dentro de uno o varios tipos de violencia:*

1) ***Umbral de peso:*** Esta regla compara los términos de cada noticia con los términos de los diccionarios para encontrar las coincidencias entre ellos. Habiendo identificado los términos que coinciden, se debe obtener la puntuación final del documento a este diccionario, la puntuación será el resultado de sumar los pesos de todas las coincidencias. Para asignarles pesos a los términos de los diccionarios se debe tener en cuenta la magnitud e impacto que refleje dicho término, en el caso del diccionario de violencia física aquellos términos que representen la muerte de la víctima serán asignados con los pesos más altos, por otro lado, en el diccionario de violencia psicológica todos los términos que reflejen discriminación hacia la víctima serán asignados con los pesos más altos, en el caso del diccionario de violencia sexual todos los términos hacen referencia a encuentros sexuales forzados, por lo que todos llevan el mismo valor de peso. Como mencionamos antes: *la probabilidad de que un tipo de violencia ocurra tiene un valor entre [0,1], un valor de 0 corresponde a que el tipo de violencia no ocurre y un valor de 1 corresponde a que el tipo de violencia ocurrirá.* Entonces se consideró que el umbral óptimo para que ocurra un tipo de violencia es el 25% del valor de ocurrencia (1), dicho valor está distribuido mediante pesos entre todos los términos del diccionario. Como resultado, aquellos documentos que cumplan con el umbral de existencia a un determinado diccionario, serán etiquetados con el tipo de violencia correspondiente al diccionario [119].

En la Figura 33 se muestra un ejemplo donde un documento obtiene 3 coincidencias con en el diccionario de violencia física, una coincidencia con el diccionario psicológico y ninguna coincidencia con el diccionario sexual. La suma de los pesos de estas coincidencias en el caso de la violencia física es superior e igual al umbral de ocurrencia 0.25, por lo tanto, este documento se etiqueta como un caso de violencia física (F). Caso

contario los pesos de las coincidencias de los diccionarios violencia psicológica y sexual no cumplieron el umbral sugerido para el etiquetado [119].

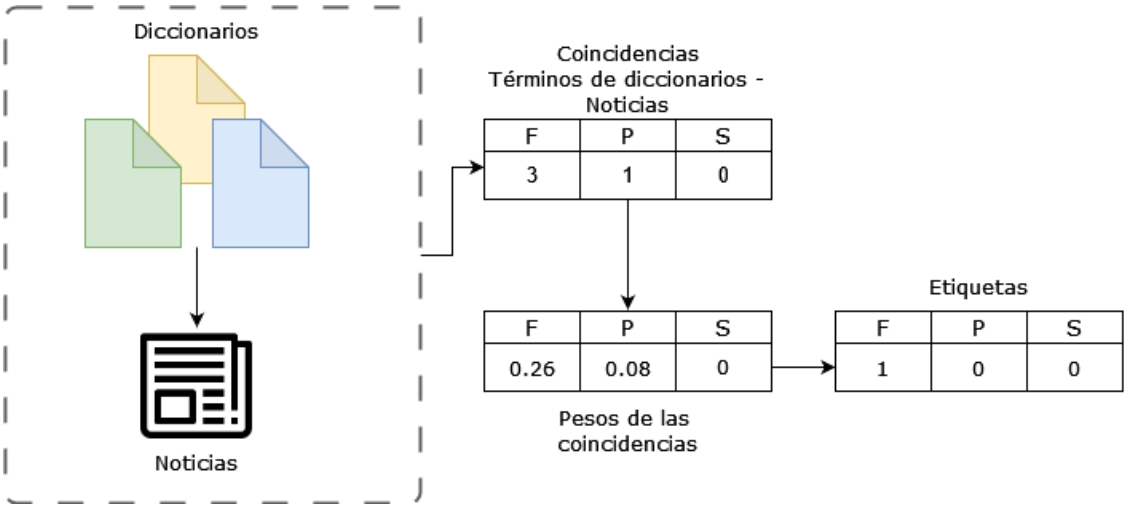

Figura 33. Etiquetado por umbral de peso.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

2) **Número de coincidencias:** Para etiquetar un documento con un tipo de violencia se siguió el soporte de un experto y se estableció: *si un documento presenta al menos 3 coincidencias con un diccionario será suficiente para considerar que existe un tipo de violencia específico.* Por lo tanto, un documento con al menos 3 coincidencias con un diccionario será identificado y etiquetado con el tipo de violencia al que pertenece ese diccionario [119].

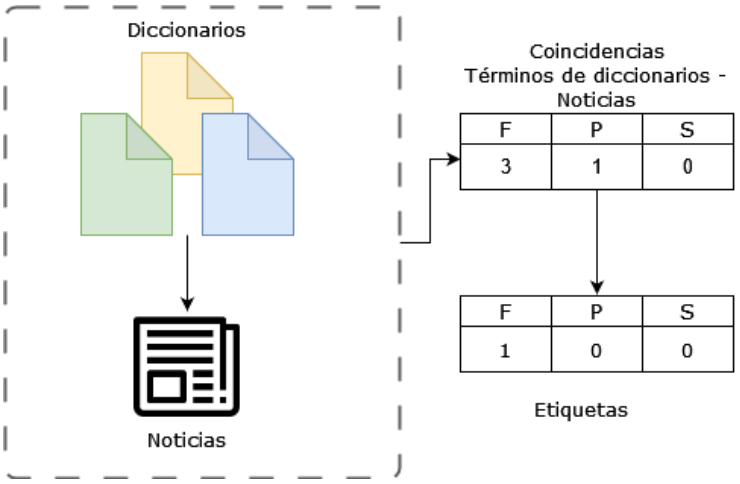

Figura 34. Etiquetado por número de coincidencias.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la Figura 34 se ilustra un ejemplo en el que un documento obtuvo 3 coincidencias con el diccionario Físico (F) y una coincidencia con el diccionario Psicológico (P). En este caso el documento fue etiquetado como F (tipo de violencia física) ya que se cumple la regla al existir al menos 3 coincidencias con cualquier diccionario [119].

3) **Términos fuertes:** Por último, los diccionarios están conformados por varios términos. Acorde a la opinión del experto: en el caso del diccionario de violencia física, cualquier acto violento que describa la muerte de la víctima será razón suficiente para etiquetarlo con ese tipo de violencia. De igual forma, en el caso del diccionario violencia psicológica, aquellos términos que representen discriminación hacia la víctima serán considerados como fuertes. Por el contrario, en el diccionario de violencia sexual, todos los términos son considerados como fuertes debido a la naturaleza de forzar a una persona a tener relaciones íntimas. En consecuencia, todo documento que presente al menos un término coincidente con alguno de los términos fuertes de algún diccionario será etiquetado con la clase de dicho diccionario [119].

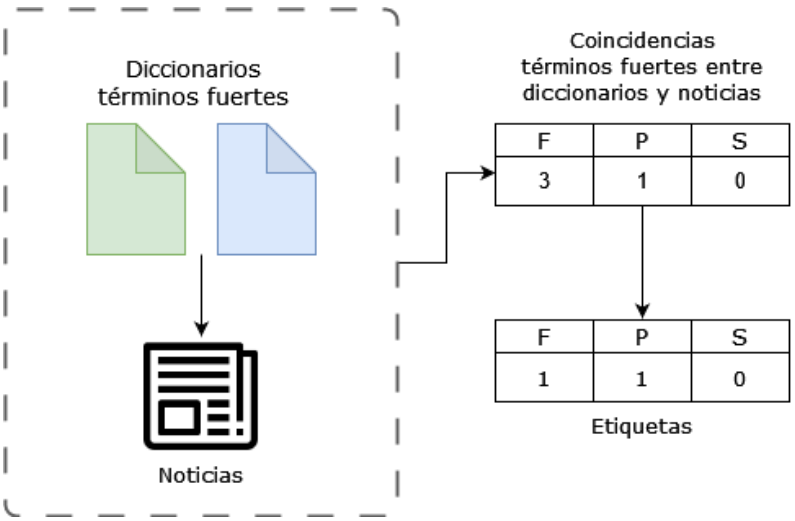

Figura 35. Etiquetado por términos fuertes.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la Figura 35 se muestra un ejemplo de un documento que obtuvo 2 coincidencias con los términos más relevantes del diccionario físico (F), y 1 coincidencia con el diccionario psicológico (P). En este caso, el documento fue etiquetado como (violencia Física y Psicológica) F y P, ya que en ambos casos se cumplía la condición de tener al menos un término de alta relevancia [119].

iii. ETIQUETADO BASADO EN PESOS, COINCIDENCIAS Y RELACIONES ENTRE LOS TIPOS DE VIOLENCIA.

Se debe destacar que el proceso de etiquetado basado en las relaciones entre los diferentes tipos de violencia se deriva inicialmente de la detección de los tipos de violencia F (física), P (psicológica) y S (sexual) para lo cual, primeramente, se aplicaron las reglas explicadas en el apartado anterior b) referente al *"Etiquetado basado en coincidencias entre los términos de los diccionarios y los términos de los documentos"* [119].

Además del método explicado anteriormente, se llevó a cabo una segunda metodología de etiquetado. En esta metodología se consideró que los tipos de violencia pueden estar relacionados entre sí. Aplicamos dos metodologías de etiquetado, con el objetivo final, de evaluar el rendimiento de los clasificadores en cada enfoque al clasificar texto en los diferentes tipos de violencia que sufre la mujer [119].

Para establecer la relación existente entre cada tipo de violencia se siguieron los siguientes conceptos: Sanmartín [147] menciona que la violencia sexual (S) es cualquier tipo de conducta en la que se obliga a un individuo para obtener estimulación o gratificación sexual. Rodríguez López et al. [148] describe que las víctimas de abuso sexual, especialmente si son niños, presentarán secuelas psicológicas, trastornos depresivos y bipolares, ansiedad, estrés y conductas autodestructivas. Por ello, el tipo de violencia sexual está compuesta e involucra daño físico y psicológico. Así mismo, en otra investigación realizada a diferentes tipos de víctimas estas demostraron que la violencia física, psicológica y sexual provocan una serie de consecuencias negativas tanto físicas como psicológicas [16]. Hernández Ramos et al. [12] indica que las agresiones físicas y/o sexuales siempre producen alguna secuela, daño o trauma psicológico. Así, destacan

que la violencia psicológica es un conjunto de actitudes y conductas donde una agresión o abuso se realiza de una forma más sutil y más difícil de detectar, evaluar y probar. En consecuencia, la violencia psicológica puede darse sin que ocurra otro tipo de maltrato (físico o sexual) [119].

A partir de los conceptos expuestos en el párrafo anterior, pudimos definir y concluir que los diferentes tipos de violencia pueden estar relacionados de la siguiente forma: cuando se detecte violencia física en un documento, se tratará también de violencia psicológica. Si un documento manifiesta violencia psicológica, entonces no se vinculará a ninguna otra violencia. Y si se detecta violencia sexual este documento estará fuertemente ligado a violencia física y psicológica. Este razonamiento queda recogido en la Tabla 2 donde F es violencia física, P es violencia psicológica y S es violencia sexual [119].

Tabla 2. Relaciones entre violencias.

Violencia Detectada	Violencia Relacionada
F	P
P	-
S	F P

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la Tabla 2 podemos ver que un documento fue inicialmente detectado y etiquetado como violencia física (F), pero basándonos en las relaciones entre las violencias este también se etiquetará con violencia psicológica (P), es decir, la etiqueta o clase final será FP. Por otro lado, si el documento hace referencia a violencia sexual (S) este involucrará violencia física (F) y violencia psicológica (P) a su vez, por lo que se clasificará como FPS [119].

Como podemos observar la Figura 36 recoge los pasos que hemos seguido hasta llegar a etiquetar el conjunto de documentos.

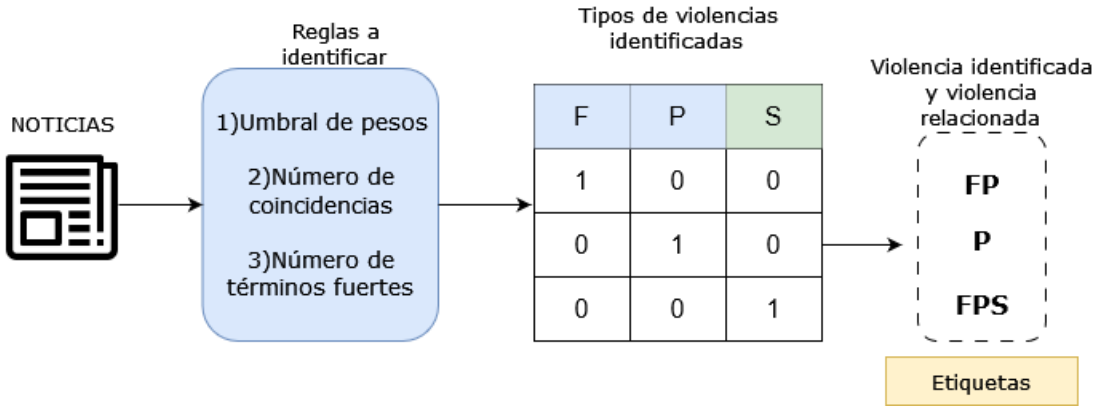

Figura 36. Etiquetado por relaciones de violencias.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

B. TRANSFORMACIÓN DEL TEXTO A FORMATO NÚMERICO

La matriz TF-IDF es ampliamente utilizada en el campo del Procesamiento de Lenguaje Natural esta herramienta es una medida estadística que evalúa la relevancia de una palabra en un documento de una colección de documentos. En esta investigación se aplicó TF-IDF con el método n-gram para transformar la colección de documentos (textos) en un documento de espacio vectorial que contenga las palabras con su respectivo número de ocurrencias e importancia dentro de la colección de documentos, para trasformar el texto se aplicó el siguiente enfoque [119]:

- Aplicando la matriz TF-IDF y mediante la configuración de sus parámetros se representó a las palabras en una secuencia de longitud [1,2]. Esto significa que las palabras fueron representadas en forma de unigramas y bigramas. Además, se descartaron aquellas palabras con una frecuencia en los documentos mayor a 0.85 porque al ser muy frecuentes no aportan valor al texto. Asimismo, se descartaron las palabras con una frecuencia en los documentos menor a 0.15 este valor expresa que los términos aparecen rara vez, igual que en el umbral anterior, no aportan utilidad al texto. Para realizar el filtrado de las palabras se aplicaron los parámetros de la matrix TF-IDF; $mín\_df = 0.15$, de esta forma, se ignoró los

términos que tengan una frecuencia de documentos estrictamente inferior al umbral dado y con el parámetro *máx_df = 0.85* se ignoró los términos que tienen una frecuencia de documentos estrictamente superior al umbral dado [119].

C. SELECCIÓN DE CARACTERÍSTICAS

La Selección de Características es un método para reducir el número de variables de entrada en un modelo tomando solo aquellas que son de relevancia. Es también conocida como selección de variables, es una tarea relevante que consiste en seleccionar un subconjunto de rasgos importantes para su uso en la construcción de los modelos. Su aplicación puede afectar directamente la dimensionalidad del conjunto de datos, de igual forma, los resultados de los modelos ya sea mejorando su valor de precisión o aumentando el rendimiento. En esta investigación previo a la clasificación de los documentos para seleccionar las mejores características para el modelado se tomó el siguiente enfoque:

- Aplicando Chi-cuadrado ($\chi 2$) se seleccionaron las 1000 características con los mejores valores de Chi-cuadrado ($\chi 2$). Se seleccionó esta cantidad de características debido a que los modelos de clasificación presentaron mejor rendimiento en la medida "Accuracy" trabajando con esta cantidad de características [119].

D. EQUILIBRIO DE CLASES

El balanceo o equilibrio de clases se realiza para evitar que en la etapa de modelado las predicciones sean erróneas beneficiando a las clases mayoritarias. Después del proceso de etiquetado del conjunto de entrenamiento, aplicando la metodología: *"b) etiquetado basado en coincidencias entre los términos de los diccionarios y los términos de los documentos"* se generaron 7 clases distintas a predecir; [F,FP,FS,FPS,P,S,PS]. En la Figura 37 se puede observar que la distribución de las clases es desequilibrada y claramente se evidencia que la clase mayoritaria es el tipo de violencia física (F) [119].

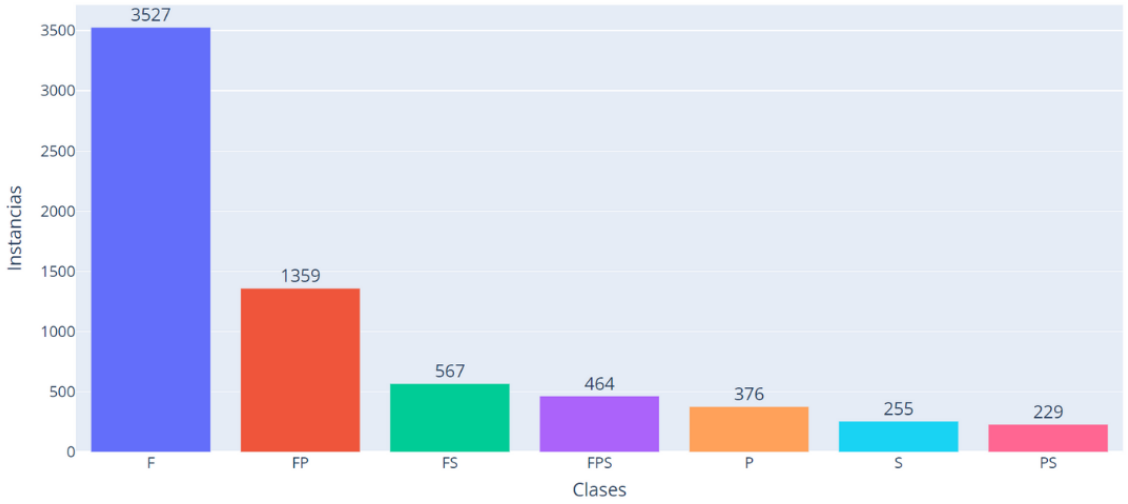

Figura 37. Distribución clases método de entrenamiento de conciencias entre términos.

Fuente: Elaboración propia.

Posteriormente, aplicando el paso: *"c) etiquetado basado en pesos, coincidencias y relaciones entre los tipos de violencia"* sobre el mismo conjunto de datos se obtuvieron 3 clases a predecir; [FP,FPS,P]. En la Figura 38 se puede observar que nuevamente la distribución es desequilibrada y de igual forma existe una clase mayoritaria la cuál es el tipo de violencia física asociada con la violencia psicológica (FP) [119].

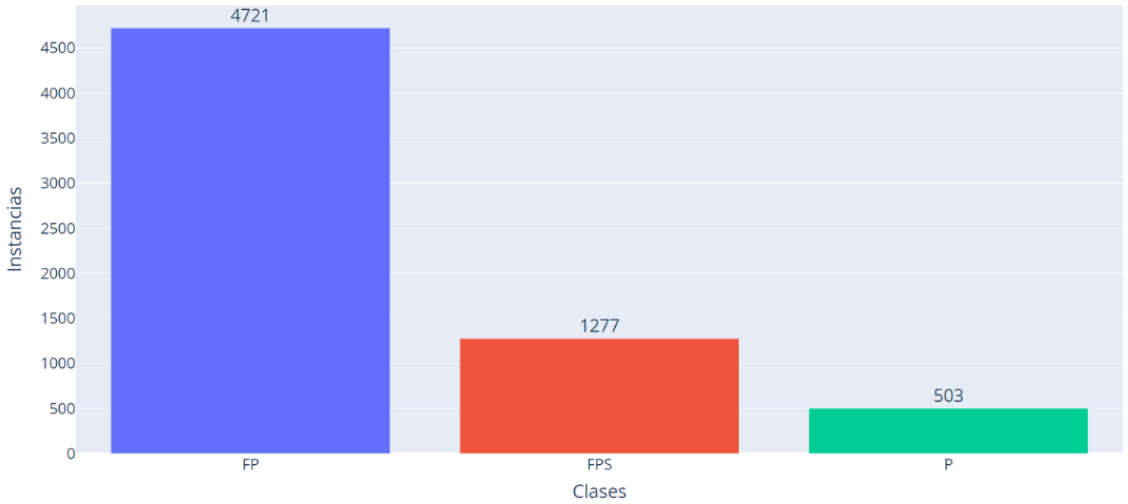

Figura 38. Distribución clases método de entrenamiento relaciones entre violencias.

Fuente: Elaboración propia.

Cómo podemos ver en la Figura 37 y la Figura 38 este desequilibrio de las clases puede afectar el rendimiento del clasificador perjudicando a las clases minoritarias. Para corregir este problema se aplicó la técnica de balanceo de clases SMOTE (ver sección B.4.). Esta técnica se encarga de crear muestras sintéticas en aquellas clases minoritarias que poseen

pocas instancias, de esta forma, las clases fueron equilibradas y se evitó que el modelo fuera sesgado hacia las clases mayoritarias [119].

3.1.3 CLASIFICACIÓN DE TIPOS DE VIOLENCIA CONTRA LA MUJER

Con respecto al proceso de encontrar un modelo de clasificación para los documentos se aplicaron 6 algoritmos. Los modelos aplicados en esta investigación fueron: Bosques Aleatorios (RF), Soporte de vectores de máquina (SVM), Arboles de Decisión (DT), Xgboost, Redes Bayesianas (NB) y Redes Neuronales Artificiales (ANN). Para obtener los mejores hiperparámetros de cada modelo se utilizó la función (GridSearchCV). Con esta función se hizo una búsqueda exhaustiva sobre los parámetros específicos de cada uno de los clasificadores. Esta función prueba todas las combinaciones posibles entre los parámetros y devuelve la combinación que obtiene la mejor puntuación [119]. Con el fin de mejorar el proceso de clasificación se realizó validación cruzada con un valor k = 10. Por otra parte, para la evaluación de los modelos se decidió usar y analizar las métricas: *Accuracy,* adicionalmente, se decidió analizar el "*F1 score",* dicha medida también ha sido aplicada para la evaluación de aplicaciones sobre texto.

Con la medida *"Accuracy"* se obtuvo la proporción que representa el desempeño del algoritmo para clasificar los datos correctamente. Se conoce que es frecuentemente usada en aplicaciones sobre texto [108] convirtiéndola en la medida con más relevancia en nuestra evaluación de los clasificadores de texto. Mediante la aplicación de "*Precision"* se midió el porcentaje entre la cantidad de muestras correctamente clasificadas en relación al total de muestras. Por último, con "*F1 score"* para describir la exactitud global de clasificación de un modelo teniendo en cuenta tanto la "precisión" (también conocida como especificidad, o tasa de verdaderos negativos) y el *"recall"* (también conocida como sensibilidad o tasa de verdaderos positivos).

3.2 IDENTIFICACIÓN DE TEMAS LATENTES SOBRE LA VCM

El modelado de temas consiste en la búsqueda de grupos de palabras (temas) a partir de documentos o una colección de documentos, dichas palabras tendrán la capacidad de representar la información contenida en el documento. En esta sección detallaremos las tareas realizadas en la fase exploratoria de los documentos, la fase del modelado para la identificación de temas latentes sobre la VCM además de la elaboración de etiquetas que representen cada tema.

3.2.1 ANÁLISIS EXPLORATORIO DEL CONJUNTO DE DOCUMENTOS

A partir de la colección de documentos recuperados mediante la aplicación de técnicas de Raspado Web, se decidió realizar un análisis exploratorio de los datos, posteriormente se realizó el tratamiento de los datos (este proceso ya fue detallado en la sección procesamiento de texto citado anteriormente), en este caso para poder mejorar la calidad de los resultados no se realizó "stemming" debido a que al contener solo las raíces de las palabras dificultaba la interpretación de los términos dentro de los temas. En este análisis se generaron los bigramas y trigramas más frecuentes de modo que podamos observar el comportamiento del conjunto de datos.

3.2.2 GENERACIÓN DE TEMAS SOBRE LA VCM

Entre las técnicas de Modelado de Temas encontramos a Latent Dirichlet Allocation (LDA), es un modelo popular usado en el modelado de temas, surgió como uno de los métodos elegidos para el análisis de documentos grandes [83]. Debido a la popularidad que ha alcanzado el método LDA para analizar grandes colecciones de documentos se decidió aplicar este método en nuestra investigación. Un parámetro importante requerido en este método es el número óptimo de temas a generar al que nombraremos como k. Para determinar el número k de temas a generar se aplicaron las medidas "log-likelihood y perplexity". En nuestro problema de VCM, estas medidas mostraron que el número

óptimo de temas a generar es de 16 tópicos con un valor de 55 en el número de iteraciones [119].

3.2.3 ETIQUETADO DE TEMAS Y DETECCIÓN DE NOTICIAS RELEVANTES

Con LDA se generaron los temas latentes junto con sus 15 palabras más frecuentes. Basados en la probabilidad de distribución de los documentos para cada tema se tomaron los 20 documentos con los valores de probabilidad más altos asociados a cada tema. A partir de ellos calculamos las 15 palabras más frecuentes y mediante un análisis de estas palabras construimos las etiquetas de identificación de los temas [149]. Queremos hacer notar que hubo temas que no se pudieron etiquetar. Esto se debe a que estos temas contaban con una distribución de probabilidad muy baja. Y como consecuencia, no pudimos obtener los 20 documentos más importantes que representarían su etiqueta [119].

Adicionalmente, se detectaron las dos noticias más relevantes por tema. Para la identificación de estas noticias (documentos) nosotros tomamos los dos valores más altos de la distribución probabilística de documentos por temas generados a través de LDA. Debido a la extensa dimensión que pueden presentar los documentos se tomó como referencia solo el título de los mismos para dar una idea general del contenido. Las etiquetas creadas y asignadas a los temas son capaces de representar claramente el contenido principal de la distribución probabilística de documentos dentro de ese tema. Estas etiquetas son una herramienta útil para la búsqueda de temas y así capturar la idea principal de los documentos dentro de la misma distribución [119]. Los detalles asociados a los resultados obtenidos pueden ser consultados en el capítulo siguiente dedicado a la Experimentación.

3.3 BUSQUEDA DE PATRONES ACERCA DE LA VCM

La búsqueda de patrones se realiza mediante la aplicación de minería de reglas de asociación. Esta es una poderosa técnica de minería de datos que permite determinar la correlación entre enormes colecciones de datos. Aplicando reglas de asociación podemos identificar relaciones entre palabras claves y así analizar su comportamiento [150]. En

esta sección detallaremos las tareas realizadas en las fases de procesamiento, construcción de características, identificación de características, la detección y generación de reglas de asociación.

3.3.1 TRATAMIENTO DEL TEXTO PARA EXTRAER REGLAS DE ASOCIACIÓN

El tratamiento del texto es importante para eliminar cualquier contenido que pueda interferir en el buen desempeño de un modelo al momento de generar reglas de asociación. Previo a la generación de las reglas de asociación se realizaron las siguientes tareas para mejorar la calidad del texto: limpieza del texto (eliminación de enlaces, números), tokenización (separar las palabras por tokens), eliminación de palabras vacías (eliminación de las palabras que no aportan valor al texto) y la obtención de las raíces de las palabras (stemming). No describiremos ampliamente estas tareas debido a que ya fueron citadas y explicadas en la sección "procesamiento del texto". Adicionalmente, para extraer reglas de asociación sobre la VCM se realizó una detección de palabras únicas.

Después de haber eliminado las palabras vacías los documentos seguían conteniendo una alta dimensionalidad. Como forma de solventar este problema, se analizó cada documento para detectar y eliminar todas las palabras repetidas dentro de éste. La Tabla 3 se ilustra como al eliminar las palabras repetidas se logra reducir la dimensionalidad del texto. Por ejemplo, las palabras *hombre* y *esposa* están presentes más de una vez en el mismo documento, por lo que, se procede a eliminar las repeticiones. Como hemos comentado antes, esto ayuda a reducir la dimensión de los documentos. Esta acción se realizó en cada documento de la colección.

Tabla 3. Representación de eliminación de palabras repetidas.

Eliminación palabras que aparecen más de una vez en el documento	
frase	ciudad méxico hombre asesinó esposa medio hombre acabar esposa hijos…
Palabras únicas	ciudad méxico hombre asesinó esposa medio acabar hijos…

Fuente: Elaboración propia.

3.3.2 CONSTRUCCIÓN E IDENTIFICACIÓN DE CARACTERÍSTICAS SOBRE LA VCM

Como comentamos en el capítulo relacionado con el marco teórico. La violencia contra la mujer es un problema de carácter social y su estudio involucra el análisis de varias características presentes en los actos de violencia. En esta investigación se consideró importante analizar las siguientes características: *tipo de víctimas, tipo de agresor, tipo de arma, motivos del ataque, tipo de violencia física, tipo de violencia sexual, tipo de violencia psicológica, tipo de heridas, además, determinar si el agresor cometió suicidio después de los hechos de violencia, y si, como resultado de la agresión la víctima perdió la vida.* Para poder identificar las características antes mencionadas, se le asignó a cada una de ellas un grupo de términos que la representan. Por ejemplo, la característica *tipo de agresor*, está compuesta por los términos: *cuñado, padre, progenitor, padrastro, amigo, abuelo, primo y hermano*.

Dicho lo anterior, para determinar los términos que conforman cada tipo de las característica a estudiar se utilizó tesauros sobre la VCM [11]. De igual forma, algunos términos se extrajeron de la frecuencia de palabras obtenida de la colección de documentos en el proceso de procesamiento del texto. Adicionalmente, se decidió *determinar si en los documentos que describen un acto de VCM el agresor se suicidó después de la agresión, asimismo, si el documento refiere a un caso de asesinato u homicidio de la víctima.*

[11] https://e-archivo.uc3m.es/bitstream/handle/10016/21533/tesauro_IEG_2015.pdf?sequence=1&isAllowed=y

En la tarea de identificación de características tanto los términos de los documentos como los términos de las diferentes características están representados por sus raíces. Como ejemplo en la Figura 39 se puede observar la representación de algunos términos de la característica "tipo de víctima y violencia física" junto a sus raíces, este proceso de transformación se realizó a todos los términos de las diferentes características.

Representación extracción de las raíces de los términos (STEMMING)									
Términos	esposa	degollar	ahorcar	niña	adolescente	estudiante	novia	torturar	mutilar
Raíces	espos	degoll	ahorc	niñ	adolescent	estudiant	novi	tortur	mutil

Figura 39. Representación términos de la característica tipo de víctima a sus raíces.
Fuente: Elaboración propia.

Cabe señalar que, en la minería de reglas de asociación los documentos pasarán a ser *transacciones,* las mismas que, estarán constituidas por las palabras que conformaban el documento. A partir de lo anterior, el siguiente paso consistió en determinar las coincidencias entre los términos de las *transacciones* y los *términos de las características* a estudiar. Como resultado, las coincidencias encontradas conformaron la lista de *itemsets* de cada *transacción*. En este caso, los *itemsets* están representados por abreviaciones, debido a que, generar reglas de asociación donde sus ítems son raíces puede resultar difícil de entender e interpretar. Dicho proceso se puede observar en la Figura 40. Para brindar una interpretación más clara de las reglas generadas, los términos que conforman los *tipos de características* fueron representados por una abreviación. Los tipos de características, sus términos y abreviaciones se pueden observar en las siguientes tablas:

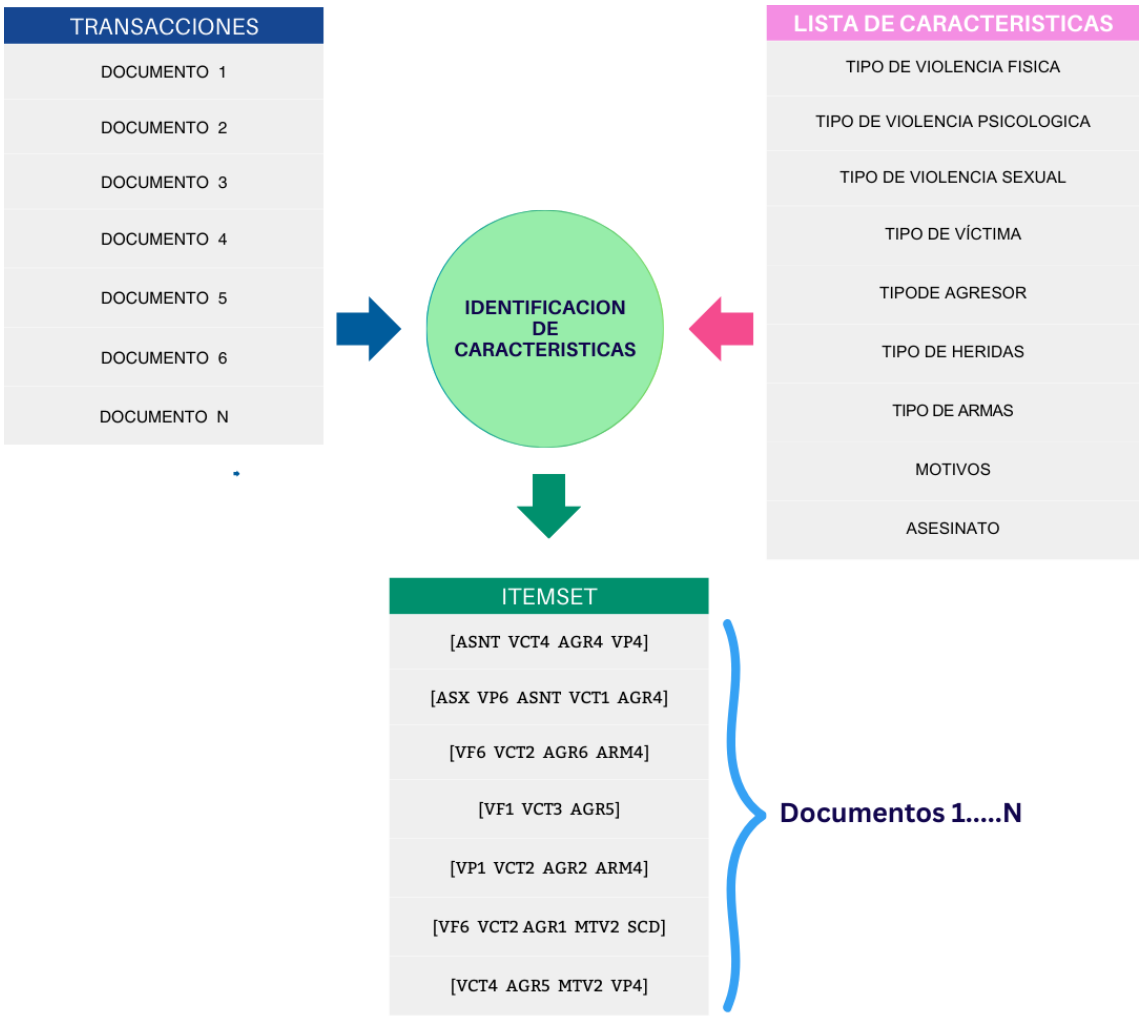

Figura 40. Representación búsqueda de características en las transacciones.

Fuente: Elaboración propia.

En la Tabla 4 se detalla la característica "tipo de violencia física", sus términos representan las diferentes formas de *"violencia física"* que una mujer puede sufrir. Aquí podemos notar que uno de sus términos es" disparar", con lo que todos los términos que se deriven de esta raíz serán etiquetados con la abreviatura "VF11". Las abreviaturas serán de ayuda para generar reglas de asociación de fácil comprensión, esta representación se lleva a cabo con todos los términos de las diferentes características. Cada término (raíz) fue representado por una abreviación distinta.

Tabla 4. Tipos de violencia física.

VIOLENCIA FÍSICA			
Términos	**Abreviación**	**Términos**	**Abreviación**
lesionar	VF1	empalar	VF14
acuchillar	VF2	mutilar	VF15
ahorcar	VF3	degollar	VF16
apuñalar	VF4	ahogar	VF17
asfixiar	VF5	amputar	VF18
cercenar	VF6	seccionar	VF19
decapitar	VF7	machetear	VF20
calcinar	VF8	balear	VF21
descuartizar	VF9	torturar	VF22
desmembrar	VF10	golpear	VF23
disparar	VF11	atacar	VF24
estrangular	VF12	machetazos	VF25
desnucar	VF13	herir	VF26

Fuente: Elaboración propia.

En Tabla 5 se muestra la característica "tipo de violencia psicológica". Esta tabla contiene los diferentes tipos de violencia psicológica junto con las abreviaciones que le asignamos a cada término (raíz) y a sus palabras derivadas previo al proceso de generación de reglas de asociación.

Tabla 5. Tipos de violencia psicológica.

VIOLENCIA PSICOLOGICA			
Términos	**Abreviación**	**Términos**	**Abreviación**
intimidar	VP1	ridiculizar	VP9
asediar	VP2	ofender	VP10
racismo	VP3	menospreciar	VP11
genocidio	VP4	amenazar	VP12
discriminar	VP5	sexismo	VP13

misoginia	VP6	machismo	VP14
humillar	VP7	degradar	VP15
insultar	VP8	secuestrar	VP16

Fuente: Elaboración propia

Finalmente en la Tabla 6 se especifica una lista que recopila otras características sobre la VCM que serán objeto de estudio. Aquí podemos ver los términos de las diferentes características, y sus respectivas abreviaciones.

Tabla 6. Lista de características violencia contra la mujer.

TIPO DE CARACTERÍSTICA	TÉRMINOS	ABREVIACIONES
Tipo de Violencia Sexual	Abuso sexual	ASX
	Violación sexual	VSX
Tipo de Víctima	Esposa	VCT1
	Ex	VCT2
	Anciana	VCT3
	Niña	VCT4
	Adolescente	VCT5
	estudiante	VCT6
	Novia	VCT7
	pareja	VCT8
	Conviviente	VCT9
	amante	VCT10
Tipo de Agresor	Cuñado	AGR1
	Padre	AGR2
	Progenitor	AGR3
	Padrastro	AGR4
	Amigo	AGR5
	abuelo	AGR6
	Primo	AGR7
	Hermano	AGR8
Tipo de Armas	martillo	ARM1
	hacha	ARM2

	revolver	ARM3
	Escopeta	ARM4
	Piedra	ARM5
	Palo	ARM6
	Cartuchera	ARM7
	Puñal	ARM8
	cuchillo	ARM9
	bate	ARM10
	Acido	ARM10
	Objeto corto punzante	ARM12
	destornillador	ARM13
	gasolina	ARM14
Tipo de Heridas	Cabeza	HRD1
	Cuello	HRD2
	brazo	HRD3
	Mano	HRD4
	Estomago	HRD5
	Abdomen	HRD6
	Muslos	HRD7
	Espalda	HRD8
	Costilla	HRD9
	Riñón	HRD10
	Hígado	HRD11
Motivos	Celos	MTV1
	Venganza	MTV2
	alcohol	MTV3
	odio	MTV4
	terrorismo	MTV5
Otros	Suicidio	SCD
	Asesinato	ASNT
	Prófugo	PRF
	Detenido	PRS

Fuente: Elaboración propia.

4 EXPERIMENTACIÓN

Este capítulo está dedicado a exponer la experimentación realizada en nuestra investigación. A continuación, detallaremos los resultados obtenidos por los modelos de aprendizaje automático en el estudio de la violencia contra la mujer (VCM) en los tres tipos de problemas planteados en la sección anterior: Clasificación de noticias, Modelado de temas y Extracción de reglas de asociación.

4.1 CLASIFICACION DE NOTICIAS EN TIPOS DE VCM

En esta investigación para clasificar las noticias en los diferentes tipos de violencia que sufre la mujer se siguió dos metodologías: "*clasificación basada en el etiquetado: coincidencias entre los términos de los diccionarios y los documentos" y "clasificación basada en el etiquetado: pesos, coincidencias y relaciones entre los tipos de violencia"*. A continuación, se presentan los resultados de los clasificadores en cada escenario de la experimentación.

4.1.1 CLASIFICACIÓN BASADA EN EL ETIQUETADO: COINCIDENCIAS ENTRE LOS TÉRMINOS DE LOS DICCIONARIOS Y LOS DOCUMENTOS.

En este apartado, presentaremos los resultados de los algoritmos de clasificación: Bosques Aleatorios (RF), Soporte de vectores de máquina (SVM), Arboles de Decisión (DT), Xgboost, Redes Bayesianas (NB) y Redes Neuronales Artificiales (ANN), al aplicar la metodología de etiquetado: *mediante las coincidencias entre los términos del diccionario y los términos de los documentos,* siguiendo el enfoque TF-IDF y Chi-cuadrado (X^2) para la selección de características. Estos resultados se pueden ver en la *Tabla 7*.

Entre los datos que recoge la *Tabla 7* se observa que los algoritmos con mejor desempeño fueron XGBoost y Árboles Aleatorios (RF). El modelo XGBoost generó las mejores métricas tanto al seleccionar características con TF-IDF y Chi-cuadrado (X^2). Sin

embargo, con una diferencia mínima entre ellos, se puede decir que el mejor desempeño se obtuvo mediante la selección de características con Chi-cuadrado (X^2). El modelo XGBoost en conjunto con Chi-cuadrado (X^2) obtuvieron una medida *Accuracy* de 0.9004546 y un valor *F1* de 0.7029262. Estos valores muestran que un alto porcentaje de los documentos fueron clasificados en la categoría correcta.

Por otro lado, Bosque Aleatorio (RF) obtuvo un valor *Accuracy* de 0.8461354, y una medida *F1* 0.6844626. Sin embargo, dichos valores fueron ampliamente mejorados por XGBoost. Los resultados muestran que XGBoost fue más eficiente al clasificar noticias, demostrando un alto desempeño y resultados más sólidos. Es interesante mencionar que, en este enfoque, aunque las redes neuronales son uno de los modelos más populares actualmente, presentaron un bajo desempeño al clasificar noticias en tipos de violencia contra la mujer.

Tabla 7. Métricas Clasificación en pesos, coincidencias.

Metodología						
Modelo	TF-IDF			Chi-cuadrado (X^2)		
	ACC	**F1**	**PRECISION**	**ACC**	**F1**	**PRECISION**
RF	0.7848767	0.5182345	0.7703329	0.8461354	0.5945612	0.7791192
SVM	0.7147642	0.3594943	0.7521954	0.7415649	0.4136052	0.7153071
NB	0.5577889	0.3653707	0.3620269	0.5664034	0.3858983	0.3663944
DT	0.5029911	0.2824870	0.2818853	0.6075616	0.2778052	0.3343335
XGB	0.8944723	0.6844626	0.6675434	**[0.9004546**	**0.7029262**	**0.7263224]**
ANN	0.5822557	0.4339038	0.5393884	0.6756002	0.4042955	0.5228558

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

4.1.2 CLASIFICACION BASADA EN EL ETIQUETADO: PESOS, COINCIDENCIAS Y RELACIONES ENTRE LOS TIPOS DE VIOLENCIA.

En esta sección reunimos los resultados obtenidos al aplicar *la metodología de etiquetado basada pesos, coincidencias y relaciones entre violencias*. Los resultados de la clasificación el enfoque TF-IDF y Chi-cuadrado (X^2) se pueden ver en la *Tabla 8*.

Tabla 8. Métricas Clasificación en pesos, coincidencias y relaciones entre violencias.

Metodología						
Modelo	TF-IDF			Chi-cuadrado (X^2)		
	Accuracy	F1	Precision	Accuracy	F1	Precision
RF	0.9498252	0.6740781	0.9180883	0.9603095	0.7182750	0.9160554
SVM	0.8093859	0.5592269	0.6037201	0.8148776	0.5703648	0.5984997
NB	0.7723414	0.5698878	0.5589073	0.7553669	0.5420144	0.5161359
DT	0.7126809	0.4662738	0.4519700	0.7406390	0.4732176	0.4535130
XGB	0.9827758	0.8918279	0.8466973	**[0.984523**	**0.8939695**	**0.968133]**
ANN	0.9382644	0.6773182	0.8769945	0.9376820	0.7162896	0.7465144

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

En la *Tabla 8* se detallan los resultados del proceso de clasificación. Podemos observar que en este escenario varios clasificadores obtuvieron un buen desempeño. XGBoost generó las métricas más altas usando el método de selección de características Chi-cuadrado (X^2), con una *accuracy* de 0.9845232 y una puntuación *F1* de 0.8939695. Seguido por Bosque Aleatorio (RF) con una a*ccuracy* de 0.9603095 y la mejor puntuación *F1* de 0.7182750. Además, ANN con un *accuracy* de 0.9376820 y una puntuación *F1* de 0.7162896.

A partir de estos resultados, podemos considerar a XGBoost como el mejor algoritmo de clasificación utilizando la técnica de selección de características Chi-cuadrado (X^2). Como consecuencia, este es el modelo más eficiente para la clasificación de textos. Además, cabe mencionar como muestran los resultados, la clasificación realizada seleccionando características con Chi-cuadrado (X^2) fue más precisa que la selección de características con TF-IDF.

4.1.3 COTEJO DEL RENDIMIENTO DE LAS METODOLOGÍAS DE ETIQUETADO PARA LA CLASIFICACIÓN DE NOTICIAS.

En la *Tabla 9* se muestran todos los valores de *accuracy* generados por los algoritmos en los dos métodos de etiquetado, aplicando la selección de características con Chi-cuadrado (X^2). Como puede observarse, las medidas *accuracy* y *precisión* obtenidas por los modelos en ambos escenarios. El modelo XGBoost fue el mejor clasificador, mostrando un alto desempeño, predicciones estables y medidas *accuracy* altas. Por otro lado, los

algoritmos con el más bajo desempeño fueron Redes Bayesianas (NB) y Árboles de Decisiones (DT). Estos resultados evidencian que estos dos algoritmos son los menos óptimos en el problema de clasificación de texto en los diferentes tipos de violencia.

Analizando los resultados obtenidos de las metodologías: "*Clasificación basada en pesos de coincidencias entre diccionarios y documentos" y "Clasificación basada en pesos y coincidencias y relaciones entre violencias",* se puede concluir que: el rendimiento de los clasificadores mejora al realizar una selección de características con Chi-cuadrado (X^2). Esto se podría deberse a que Chi-cuadrado (X^2) se enfoca en comprobar la asociación de las características con cada clase ayudando al algoritmo a predecir con más precisión. Al contrario, TF-IDF asigna pesos a las características basadas en la frecuencia de ellas en la colección de documentos.

Tabla 9. Medidas Accuracy sobre Chi-cuadrado (X^2) para las metodologías de clasificación.

Metodología				
Modelo	**Clasificación basada en pesos y coincidencias**		**Clasificación basada en pesos y coincidencias y relaciones entre violencias**	
	Accuracy	**Precision**	**Accuracy**	**Precision**
RF	0.8461354	0.7791192	0.9603095	0.9160554
SVM	0.7415649	0.7153071	0.8148776	0.5984997
NB	0.5664034	0.3663944	0.7553669	0.5161359
DT	0.6075616	0.3343335	0.7406390	0.4535130
XGB	**0.9004546**	**0.7263224**	**0.9845232**	**0.9681339**
ANN	0.6756002	0.5228558	0.9376820	0.7465144

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

4.2 MODELADO DE TEMAS SOBRE LA VCM

Como hemos comentado en el capítulo anterior. El modelado de temas es un método para extraer distribuciones de probabilidad de palabras ocultas, denominadas temas, de una colección de documentos o también denominada corpus de texto. El modelado temático utiliza modelos probabilísticos para descubrir la estructura semántica subyacente en la colección de documentos, a partir de un análisis jerárquico bayesiano de los textos

originales, descubriendo patrones de uso de palabras y conectando documentos que presentan patrones similares.

A continuación, expondremos los resultados obtenidos en los procesos: análisis exploratorio de los documentos, generación de temas y sus etiquetas de identificación, por último, la distribución de documentos y las noticias más relevantes por tema.

4.2.1 ANALISIS EXPLORATORIO DE LA COLECCIÓN DE DOCUMENTOS

A partir de la colección documentos recuperados mediante la aplicación de técnicas de Raspado Web. Previo al modelado de temas se realizó un análisis exploratorio de los documentos, se obtuvieron los bigramas más populares relacionados a la violencia contra la mujer, entre ellos tenemos: redes sociales, ministerio público y abuso sexual. Así mismo, los trigramas más populares fueron: víctimas violencia género, procuradoría general justicia, mujeres víctimas violencia.

En la *Tabla 10* se detallan las combinaciones de las palabras más usadas en los documentos. En el caso del bigrama *"redes sociales"* es muy frecuente, una de las causas puede ser que, debido a su popularidad y el alto uso de estos medios de comunicación, las redes sociales se han convertido en un medio muy usado para transmitir información por parte de la sociedad donde se ha vuelto común difundir casos de violencia contra la mujer o incluso algún tipo de reclamo en contra de la vulnerabilidad de las mujeres.

Tabla 10. Bigramas y trigramas frecuentes.

BIGRAMAS	FRECUENCIA	TRIGRAMAS	FRECUENCIA
Redes sociales	677	víctimas violencia género	188
Ministerio publico	607	procuraduría general justicia	159
Abuso sexual	582	mujeres víctimas violencia	125
Ex pareja	585	erradicar violencia mujeres	112
prisión preventiva	528	casos violencia género	107
Medicina legal	494	código orgánico integral	106
Fiscalía general	447	observatorio nacional feminicidio	86
Ciudad México	394	agente ministerio público	85
Juan Carlos	362	fiscalía general justicia	78
Menores edad	284	oficina violencia doméstica	68
Uribe noguera	339	denuncias violencia género	64
Trata personas	209	condenado prisión perpetua	59

Arma blanca	303	vida libre violencia	58
Arma fuego	240	heridas arma blanca	55
Homicidio agravado	207	prisión preventiva oficiosa	53

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

Además, se calcularon las palabras más frecuentes en la colección de documentos (unigramas). En la *Figura 41* se muestran las 300 palabras más frecuentes en la colección de documentos. El tamaño de la fuente refleja la frecuencia de la palabra en el conjunto de documento. En este caso se observa que *feminicidio* es la palabra con la fuente más grande representando que es la palabra más frecuente seguida de las palabras *asesinar, crimen u homicidio*, estas palabras claramente describen una acción violenta que acaba con la vida de la víctima. Por otro lado, las palabras *casa* y *pareja* están también entre las más frecuentes lo cual puede sugerir que muchos de los casos de violencia ocurren en casa y el agresor es la pareja de la víctima, convirtiéndose en violencia doméstica o familiar.

Figura 41. Palabras frecuentes.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

4.2.2 GENERACIÓN DE TEMAS Y SUS ETIQUETAS

Como comentamos, Latent Dirichlet Allocation (LDA) es un modelo popular usado en el modelado de temas, surgió como uno de los métodos elegidos para el análisis de documentos grandes. LDA permite generar temas a partir de la frecuencia de palabras de un conjunto de documentos. En esta sección detallaremos los temas generados mediante la aplicación de LDA.

En la *Tabla 12* se detalla los temas generados con LDA junto con sus 15 palabras más frecuentes. Mediante la distribución probabilística de los documentos a los temas generados, se tomaron los 20 documentos con los valores de probabilidad más altos para cada tema. A partir de ellos, calculamos las 15 palabras más frecuentes y mediante un análisis de estas palabras construimos las etiquetas de identificación para cada tema. Así mismo, en esta tabla se observa que los temas 6 y 13 no tienen etiquetas asignadas. Esto se debe a que estos temas estaban representados por una baja distribución de probabilidad y como consecuencia no pudimos obtener los 20 documentos más importantes que se utilizarían para generar la etiqueta.

Además, examinando la *Tabla 11* los términos asociados al tema 8 hacen referencia a situaciones de protestas, en concreto, aquí obtuvimos la etiqueta: «*manifestaciones sobre la VCM en redes sociales*». Esta etiqueta expresa que hay personas reclamando o manifestándose por la violencia que sufren las mujeres. Por otro lado, en el tema 12, sus términos reflejan que hay hechos de violencia dentro del seno familiar, aquí se obtuvo la etiqueta «*pareja homicida*». En el tema 4 sus términos detallan que entre las víctimas de violencia encontramos niñas, esto evidencia que las niñas se han convertido en un grupo vulnerable que sufren violencia física, psicológica y sexual, obteniendo la etiqueta: «*violencia contra niñas*». Finalmente, los términos del tema 0 reflejan que la sociedad lucha por obtener leyes que protejan a las mujeres, la etiqueta para este tema es: «*petición de leyes o políticas de protección e igualdad*». Los datos obtenidos expresan que las redes sociales son un medio de comunicación ampliamente usado para protestar contra la *VCM*. También, podemos mencionar que los hechos de violencia están sucediendo por personas

del núcleo familiar, siendo el esposo el agresor más común, dicha violencia está catalogada como violencia doméstica al ocurrir en el seno familiar.

Tabla 11. Temas y sus etiquetas.

TEMA	PALABRAS CLAVES	ETIQUETA
0	sistema asesinadas congreso sociales datos marcha atención política situación presidente protección denuncias mayor seguridad ley	Petición de leyes o políticas de protección e igualdad
1	mexicano asesinato transporte veracruz estudiante estatal ricardo mara fiscalía capitalina protocolo informó joven universidad seguridad	Violencia contra estudiantes
2	niña pena letal firmada florida autoridades asesinato penitenciarias mandato sometido condenado Wayne reo ejecutaron décima	Ejecución de agresor
3	soldados nezahualcóyotl lupita rojas carlos mosquera arana guerrera viudo arenas lilia buga yadira oblitas humala	----
4	sexual hechos general acusado niña autoridades tribunal cárcel muerte delitos juicio proceso pena audiencia penal	Violencia contra niñas
5	juzgado piura público perú agresor prisión ayacucho sentencia adriano judicial tentativa arlette lima contreras pozo	Detención de agresor
6	lópez masacre agravado suministro matar mercado drogas estupefacientes suministro anestesista sabrina maguna romina belén torres	violencia por injerta de estupefacientes
7	catarina damián muerte susana torres identidad asesinato jimmy activista región lesbiana gay comunidad maría lgbt	Asesinato por motivos de genero
8	sociedad periodista realidad voz obra hijos libros personas redes miedo verdad gda trabajo hablar amor	manifestaciones sobre la VCM en redes sociales
9	homicidio migrantes juan policías venezolana ojeda ciudadanos ecuatoriana carolina quito maduro moreno venezolanos ibarra diana	Violencia de inmigrantes en ecuador
10	delictiva extorsión semáforo lesiones habitantes narcomenudeo delictivo carranza secuestro homicidio incremento tasa robo guerrero homicidios	Incremento tasa homicidios
11	rosita herrera católica trujillo hermana comunidad rocha ángela miembros nicaragua muerte dios vilma pastor iglesia	Muerte por motivos religiosos
12	sospechoso barrio noche agentes asesinato hospital agrabado fiscal familiares ex vivienda muerte detenido joven pareja	Pareja homocida
13	informativo vídeos vih consentida consideraciones apoyaba reafirmó sida contagio violaban quiosco amarró mandos atribuir acostarse	-----
14	martín juanita quemaduras carlos vizcarra gasolina vulnerables poblaciones miraflores prendió fuego eyvi cajamarca lima perú	Detención de aggressor que prendió fuego a mujer
15	futbolistas gerais equipo hondureño hermida raquel boa minas asesinato jugador selección futbolista cáceres brasileña club	Prisión jugador de futbol por asesinato

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

4.2.3 DISTRIBUCIÓN PROBABILISTICA Y NOTICIAS MAS RELEVANTES POR TEMA.

Después de la aplicación de LDA sobre nuestro conjunto de datos de 7000 noticias sobre la VCM, se obtuvo la distribución probabilística de ocurrencia por tema. En la *Tabla 12* se detallan los temas generados. Aquí se observa que la etiqueta *«manifestaciones sobre la VCM en redes sociales»* correspondiente al tema 8 posee una probabilidad de ocurrencia muy alta con un valor de 0.7421%, seguido por la etiqueta *«pareja homicida»* correspondiente al tema 12 con una probabilidad de ocurrencia de 0.3489%, de igual forma, la etiqueta *«violencia contra niñas»* correspondiente al tema 4 posee una probabilidad de ocurrencia de 0.3155%, y finalmente la etiqueta *«petición de leyes o políticas de protección e igualdad»* correspondiente al tema 0 posee una probabilidad de ocurrencia de 0.2658%. Estos datos manifiestan que a partir de nuestro conjunto de datos existe un 0.7421% de probabilidad de que una noticia haga referencia a *«manifestaciones sobre la VCM en redes sociales»*. Además, con una probabilidad de ocurrencia de 0.3894 puede verse que existen actos de violencia que están sucediendo dentro del círculo familiar. Así mismo, con una probabilidad de ocurrencia de 0.3155 podemos observar que las niñas son individuos vulnerables que están sufriendo violencia física, psicológica o sexual.

Tabla 12. Distribución probabilística de los temas.

TEMA	ETIQUETA	PROBABILIDAD DE OCURRENCIA
0	*Petición de leyes o políticas de protección e igualdad*	0.2658
4	*Violencia contra niñas*	0.3155
8	*Manifestaciones sobre la VCM en redes sociales*	0.7421
12	*Pareja homicida*	0.3894

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

Así mismo, mediante la aplicación de LDA se pudo obtener las noticias más relevantes por tema. En la *Tabla 13* se especifica las dos noticias más relevantes por tema. Para identificar estos documentos, tomamos los dos valores más altos en la distribución probabilística de los documentos por cada tema obtenida aplicando LDA. Debido al extenso contenido de los documentos hemos mostrado solo el título del documento para

dar una idea general del contenido. Como puede observarse, las etiquetas creadas y asignadas a los temas son capaces de representar claramente el contenido principal de los documentos dentro de ese tema. Estas etiquetas son una herramienta útil para encontrar temas y capturar la idea principal de los documentos.

Tabla 13. Noticias más dominantes por tema.

TEMA	NOTICIA (DOCUMENTO)
Petición de leyes o políticas de protección e igualdad	*Violación en Miramar: el caso de la niña de 14 años que presuntamente fue abusada por 5 jóvenes en un camping que conmociona a Argentina.*
	Los monstruos de Ecatepec: la pareja acusada de decenas de asesinatos de mujeres recibió una primera condena de 15 años de prisión.
Pareja homocida	*Encuentran una víctima más del atacante de Lieja.*
	Violencia machista: el caso del hombre que mató a su esposa y acabó asesinando a la abogada que lo defendió, con la que tenía una relación.
Violencia contra niñas	*Caso Yuliana Samboní: de qué acusan a los hermanos de Rafael Uribe, el acaudalado arquitecto encarcelado por el asesinato que conmocionó a Colombia.*
	Quién es Eva Analía De jesús, Higui, la argentina presa por matar al hombre que la iba a violar y por qué la apoya René Higuita.
Manifestaciones sobre la VCM en redes sociales	*Una noche fingí que estaba muerta y ahí acabó el abuso sexual": Eve Ensler, autora de "Los monólogos de la vagina", revela el acoso de su padre cuando era niña.*
	Debate sobre el aborto en Argentina: por qué las autoridades ordenaron que se le practicara una cesárea a una niña de 12 años embarazada de casi 6 meses
Violencia contra estudiantes	*Vinculan por otro feminicidio a pareja de Ecatepec.*
	Nataly Michel tenía más de cinco horas muerta cuando fue hallada.
Detención de agresor	*Emite Guerrero declaratoria de Alerta de Género en 8 municipios.*
	Luis Alberto es condenado a más de 47 años de prisión por matar a golpes a 2 mujeres en Edomex.

Fuente: Study of violence against women and its characteristics through the application of text mining techniques, (Stephanie et al., 2023).

4.3 EXTRACCIÓN DE ASOCIACIONES SOBRE LA VCM

En este apartado trataremos los resultados obtenidos al aplicar reglas de asociación. A través de ellas descubriremos las relaciones existentes entre las características dentro de un conjunto de documentos. Para poder aplicarlas, inicialmente, hay que crear una lista de "itemsets" para cada "transacción" y finalmente generar las reglas de asociación que permitirán descubrir los "ítems" que están correlacionados entre sí. Ante la amplia popularidad y uso del algoritmo "Apriori" se decidió aplicarlo en esta investigación. Esta sección detalla las reglas de asociación generadas a partir de la colección de noticias sobre la VCM.

4.3.2 GENERACIÓN DE ITEMSETS Y PATRONES DE ASOCIACIÓN

En la *Tabla 14* se puede observar los resultados generados por el algoritmo Apriori, entre ellos tenemos: los "itemsets" más frecuentes de las 7000 "transacciones". A partir de esto, se detectó que 1343 víctimas fueron abusadas sexualmente representado el 19.185% de los casos de violencia que están siendo estudiados. Por otro lado, de las 7000 transacciones que contiene la base de datos, al menos 5240 mujeres fueron víctimas de algún tipo de violencia que terminó con sus vidas, representando el 74.85% de los casos de violencia almacenados en la base de datos. Así mismo, estos datos señalan que la Violencia Física es el tipo de violencia que más persiste en los casos de violencia contra la mujer, siendo el arma ARM8 (puñal) la más utilizada. Las heridas más frecuentes causadas por la violencia física se encuentran en la espalda y riñones de la víctima posiblemente realizadas con un puñal (Arma blanca, formada por una hoja de acero corta y puntiaguda). También se detectó que en el 5240 de las transacciones se encontró la abreviatura "ASN", esto significa que el 74.85% de las transacciones la víctima fue asesinada.

Tabla 14. Itemset Frecuentes.

ITEM	FRECUENCIA
(AGR2)	1250
(AGR8)	1077
(ARM8)	466
(ASNT)	5240
(ASX)	1035
(HRD1)	798
(HRD4)	1020
(VCT1)	1050
(VCT2)	1070
(VCT4)	1638
(VCT5)	770
(VCT7)	622
(VCT8)	2501
(VF1)	636
(VF11)	618
(VF23)	1239
(VF24)	929
(VF26)	958
(VP12)	746
(VP16)	506
(VSX)	1343

Fuente: Elaboración propia.

Después del reconocimiento y la generación de la lista de itemsets para cada transacción. Se observó que en el 47.27% el agresor fue detenido, esto se debe a que la abreviatura "PRS" fue encontrada en 3309 transacciones (documentos). Además, en el 9.328% el agresor se dio a la fuga, la abreviación "PRF" fue encontrada en 653 transacciones. En el 5.814% de la colección de documentos el agresor se suicidó después de agredir y asesinar a la víctima, aquí la abreviatura "SCD" fue encontrada en 407 transacciones.

En la *Tabla 15* se detallan las reglas de asociación obtenidas mediante la aplicación del modelo A priori. Estas reglas muestran que las víctimas: VCT1 (esposa), VCT2 (ex), VCT4 (niña), VCT5 (adolescente) y VCT8 (pareja) son las víctimas más frecuentes de un asesinato (ASNT). Analizando la regla **PRS→ASNT** con un soporte de 0.395520 y una confianza de 0.795104. Estas medidas indican que: en el 79% de las transacciones donde aparece la característica PRS(prisión) también aparecerá ASNT (asesinato) y que el 39% de las transacciones incluyen ambas características PRS y ASNT. Se puede decir, que un porcentaje alto de agresores que cometen asesinato son llevados a prisión. Por otra parte; la regla **HRD1→ASNT**, con un soporte de 0.1056 y una confianza de 0.8809, estos datos muestran que: en el 88% de las "transacciones" donde aparece HRD1 (cabeza)

también aparecerá ASNT (asesinato) y en el 10% de las transacciones incluyen ambas características HRD1 y ASNT, con respecto a estas métricas podemos mencionar que en relación con la regla mencionada en muchos casos de asesinato la víctima recibió heridas en la cabeza HRD1.

En cuanto a la regla **VF26→ASNT** con un soporte de 0.115604 y una confianza de 0.8027 expresa: que en el 80% de las "transacciones" donde aparece VF26 también aparecerá ASNT y en el 11% de las transacciones se incluyen ambas características VF26 y ASNT. Lo mismo ocurre con **VF23→ASNT** con un soporte de 0.1551 y una confianza de 0.8329 expresa: que en el 83% de las transacciones donde parece VF23 también aparece la característica ASNT y en el 15% de las transacciones incluyen ambas características VF23 y ASNT. Las reglas citadas anteriormente evidencian que en los casos de violencia dónde se asesinó a la víctima existió violencia física cómo heridas y golpes.

Por otra parte, el tipo de violencia psicológica la regla **VP12→ASNT** expresa que en el 79% de las transacciones donde parece VP12 (amenazar) también aparece ASNT y en que en el 10% de las transacciones incluyen ambas características VP12 y ASNT. Esto indica que el maltrato psicológico VP12 en pocos casos la víctima es asesinada esto puede deberse a que este tipo de violencia VP12 es de tipo verbal.

A su vez, hemos encontrado la regla **AGR2→ASNT** esta indica que en el 84% de las transacciones donde parece AGR2 (padre) también aparece ASNT y que el 15% de las transacciones incluyen ambas características AGR2 y ASNT. Esta regla explica que uno de los agresores que es capaz de cometer un asesinato es el padre de la víctima.

Así mismo, en la *Tabla 15* se obtuvieron las siguientes reglas **HRD4→ASNT** y **HRD1→ASNT**. Estas indican que las víctimas de asesinato comúnmente sufren las heridas en: HRD4 (mano) y HRD1 (cabeza). Finalmente, una regla que requiere mucha atención es **AGR2→VCT4** esta sugiere que existen casos de violencia donde el agresor AGR2 (padre) en lugar de proteger y cuidar de sus hijos, agrede ya sea de forma física, psicológica o sexual a su hija (VCT4).

Tabla 15. Reglas de asociación con dos ítems.

REGLAS DE ASOCIACIÓN				
ANTECEDENTE	**CONSECUENTE**	**SUPPORT**	**CONFIDENCE**	**LIFT**
VCT2	ASNT	0.126728	0.787850	1.000149
PRS	ASNT	0.395520	0.795104	1.009357
VP12	ASNT	0.089446	0.797587	1.012509
VF26	ASNT	0.115604	0.802713	1.019017
ASX	ASNT	0.125676	0.807729	1.025384
VCT8	ASNT	0.307276	0.817273	1.037500
HRD4	ASNT	0.126578	0.825490	1.047931
VCT4	ASNT	0.204449	0.394993	1.054012
VF23	ASNT	0.155141	0.832929	1.057372
AGR8	ASNT	0.135898	0.839368	1.065549
VCT1	ASNT	0.132892	0.841904	1.068769
AGR2	ASNT	0.158749	0.844856	1.072444
VCT5	ASNT	0.098015	0.846753	1.074924
HRD1	ASNT	0.105682	0.880952	1.118338
VSX	PRS	0.116055	0.574832	1.155571
AGR2	PRS	0.109140	0.580835	1.167567
AGR8	PRS	0.095610	0.590529	1.187126
VF23	PRS	0.098917	0.608555	1.223363
ASX	PRS	0.110944	0.608695	1.223645
VF23	VCT8	0.098917	0.531073	1.412515
AGR2	VCT4	0.087191	0.463999	1.884327

Fuente: Elaboración propia.

Además, en la *Tabla 16* se observa la regla **VCT8, PRS →ASNT** posee dos antecedentes, y un consecuente, esta regla obtuvo un soporte de 0.1730 y una confianza de 0.8268 estas métricas señalan que en el 82% de las transacciones donde aparecen las características VCT8 (pareja) y PRS (prisión). También aparece la característica ASNT y en el 17% de las transacciones se incluyen las características antecedentes VCT8, PRS y consecuente ASNT. Así mismo, la regla **AGR2, PRS→ASNT** con un soporte de 0.0950 y una confianza de 0.8705 señala que en el 87% de las transacciones donde aparecen las características AGR2 y PRS como antecedente también aparecerá la característica ASNT como consecuente. Además, el 9% de las transacciones contienen los antecedentes AGR2, PRS y el consecuente ASNT. Las dos reglas mencionadas anteriormente indican que en los casos donde se comete asesinato en la víctima VCT8 y VCT4 el agresor termina en prisión PRS. En el caso de la regla **ASNT, VCT4→PRS** sus métricas indican que en el 54% de las transacciones donde aparecen las características ASNT y VCT4 (niña), también aparece la característica PRS y que el 11% de las transacciones incluyen las características antecedentes ASNT, VCT4 y el consecuente PRS, podemos expresar

que muchos casos dónde se asesinó a la víctima VCT4 el agresor fue a prisión. También, tenemos la regla **ASNT, VSX→PRS,** esta indica que en menos del 50% de las transacciones donde aparecen las características ASNT y VSX estas se relacionan a la característica PRS, y que en menos del 10% de las transacciones se incluyen las características antecedentes ASNT, VSX y el consecuente PRS, en este caso la regla explica que el agresor que cometa un abuso sexual VSX, que acabe con la muerte de la víctima, en menos del 50% de los casos el agresor va a prisión PRS.

Tabla 16. Reglas de asociación con 3 ítems.

3 ITEMS ASSOCCIATION RULES				
ANTECEDENTE	**CONSECUENTE**	**SUPPORT**	**CONFIDENCE**	**LIFT**
VCT8,VCT2	ASNT	0.088544	**0.798102**	1.013164
VCT8,PRS	ASNT	**0.173030**	**0.826867**	1.049680
ASNT, VCT4	PRS	0.110493	0.540441	1.086435
AGR2, PRS	ASNT	0.095009	**0.870523**	1.105099
ASNT, VCT8	PRS	**0.173030**	0.563111	1.132009
ASNT,VSX	PRS	0.091250	0.451973	1.142731
VF23	ASNT, PRS	0.095009	1.134500	1.278316

Fuente: Elaboración propia.

A su vez, en la *Figura 42* se muestra gráficamente las dependencias entre las variables. Se puede observar que existen más casos de asesinatos cometidos por la violencia física **VF23** (golpear), que **VF26** (herir). También, hay muchas víctimas de asesinatos **VCT8** (pareja), **VCT5** (adolescente), **VCT4** (niña), **VCT1** (esposa) y **VCT2** (ex), pero la víctima más frecuente de un asesinato es la **VCT4**. Los agresores que más han cometido delitos y han ido a prisión son el **AGR2** (padre). Además, vemos que la **VCT4** (niña) tiene una relación con el **AGR2** (padre). De igual forma, observamos que hay muchos casos de carácter sexual que terminan en asesinatos (definiendo como agresión sexual todo tipo de acto de carácter sexual que no involucre contacto físico). La violencia sexual **VSX**, por otro lado, tiene relación con los casos de asesinatos **ASNT** (definiendo con violencia sexual cualquier acto de carácter sexual que implique el uso de violencia para forzar un encuentro sexual). Vemos que las agresiones físicas que más predominan en los casos de agresión sexual se infringen en las partes de cuerpo: **HRD4** y **HRD1.**

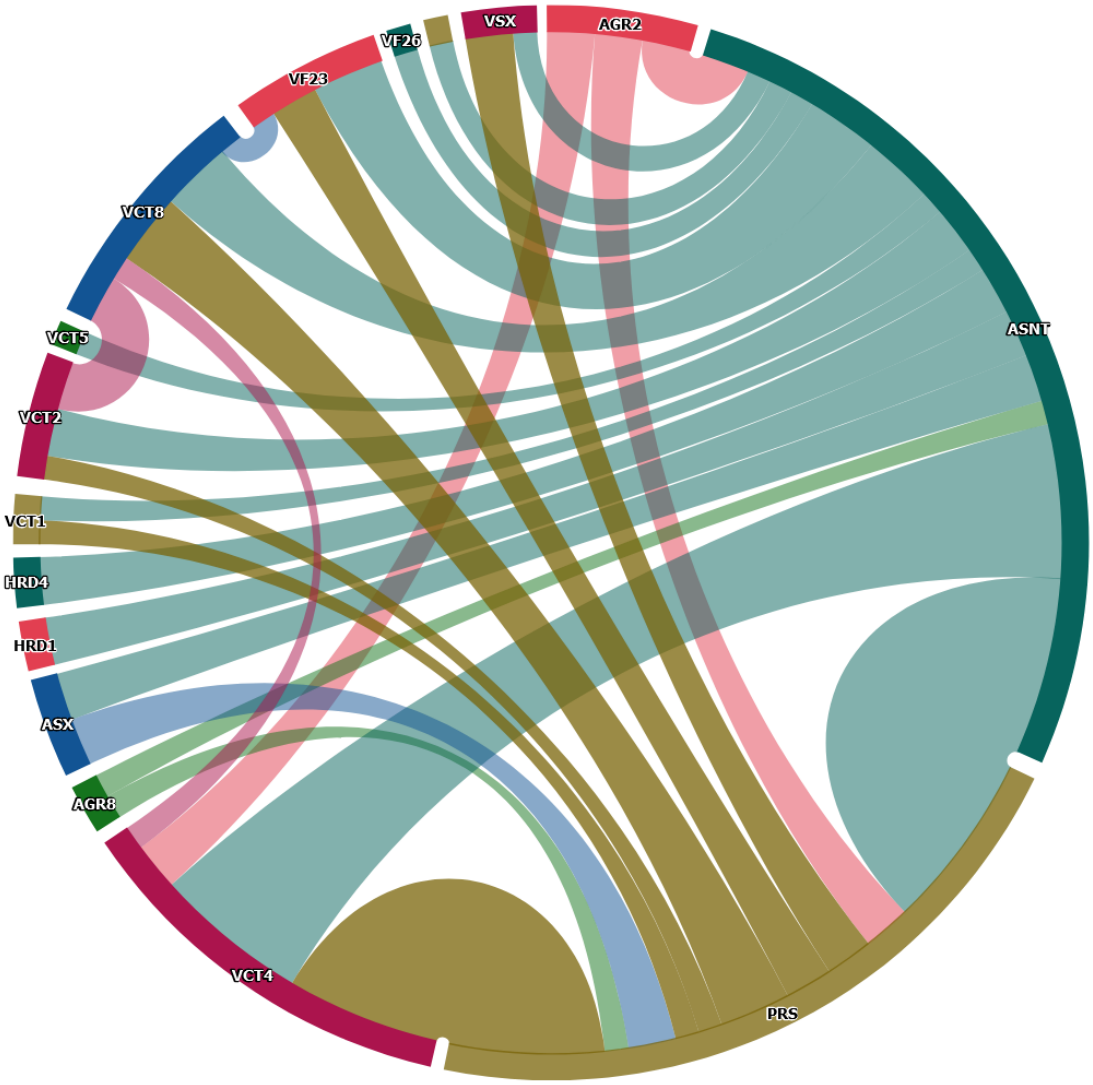

Figura 42. Dependencias entre variables.

Fuente: Elaboración propia.

5 CONCLUSIONES GENERALES Y LÍNEAS DE TRABAJO FUTURO

Una vez culminada esta investigación, en este capítulo se detallan las conclusiones que se han obtenido de las diferentes actividades llevadas a cabo en esta tesis doctoral. Cabe mencionar que este estudio se enfocó en la aplicación de la minería de texto y su ámbito en el procesamiento del lenguaje natural para el estudio de la Violencia Contra la Mujer. Además, se propondrán líneas de investigaciones futuras a considerar.

5.1 CONCLUSIONES

Durante el transcurso del tiempo la Violencia Contra la Mujer (VCM) se ha vuelto en un problema que afecta a muchas sociedades a nivel mundial. Existe un gran número de mujeres que sufren agresiones por parte de una pareja o alguna persona externa. Por lo que ha sido reconocida por la OMS como una violación de los derechos de las mujeres y su erradicación se ha convertido en un reto difícil de alcanzar. A pesar de existir leyes igualmente sanciones que condenan estos actos violentos, los casos de mujeres agredidas o asesinadas son alarmantes. Es por esto que el objetivo de esta tesis ha consistido en investigar cómo puede aplicarse la Minería de Texto, el Aprendizaje Automático y el Procesamiento de Lenguaje Natural (PLN) para realizar un estudio acerca de la VCM y así obtener conocimiento a partir de textos que provienen de noticias en diversos medios de internet. Uno de los logros importantes de esta investigación fue obtener nuevo conocimiento en el ámbito psicológico y social mediante el análisis de investigaciones y artículos sobre la VCM, con el cual se pudo adquirir una visión más amplia sobre las formas de violencia que sufren las mujeres y que en muchas sociedades no se le confiere la importancia correspondiente al sufrimiento de estas víctimas. Mediante la aplicación de diferentes técnicas de Minería de Texto se logró clasificar el texto en los tipos de violencia (física, psicológica y sexual), así mismo, se generó temas latentes y sus estructuras temáticas en publicaciones de periódicos digitales, por último, se logró extraer patrones de asociación acerca de la VCM. Con los resultados obtenidos se pretende profundizar en la problemática de la VCM, conocer las diferentes características que la

involucran, su impacto y su alcance. A continuación, se escriben los logros obtenidos en los objetivos:

- ***Objetivo 1: Explorar y analizar el proceso de Raspado Web para adquirir el conocimiento necesario que nos permita poder aplicarlo en la recolección del conjunto de datos que será estudiado.***

  En el dominio del Procesamiento de Lenguaje Natural, Análisis de Sentimientos, Minería de Texto y el aprendizaje Automático, la recuperación de los datos puede ser la fase inicial. Por lo que las técnicas de raspado web han sido ampliamente utilizadas por investigadores en diferentes campos para la recuperación de datos desde internet. En esta investigación mediante la creación de un script desarrollado en RStudio aplicando el paquete rvest, se generaron peticiones bajo un protocolo HTTP que permitió obtener una colección de 7000 documentos de texto provenientes de páginas web en formato HTML. Como conclusión, se considera que las técnicas de raspado web son una solución efectiva para lograr obtener una masiva cantidad de datos desde internet. Además, es una valiosa herramienta que posteriormente permite extraer conocimientos significativos a partir de datos no estructurados.

- ***Objetivo 2: Investigar y evaluar la aplicabilidad las diferentes técnicas de Minería de Texto para definir qué modelos o algoritmos se deben utilizar con el fin de obtener óptimos resultados.***

  El campo de la Minería de Texto cuenta con una alta variedad de algoritmos de gran potencial que han sido diseñados para cumplir una tarea específica; Clasificación, Modelado de Temas y la Extracción de Reglas de Asociación. Cabe señalar que estas tareas han logrado óptimos resultados al aplicarse sobre texto por lo que están siendo ampliamente aplicados por científicos. En esta investigación que parte de datos no estructurados se evidenció que es imprescindible la aplicación de técnicas de PLN para el tratamiento de los datos sirviendo de soporte para transformar texto no estructurado a un conjunto de datos

con estructura de forma que facilite su estudio computacional a través de diferentes algoritmos de Minería de Texto.

- ***Objetivo 3: Clasificar noticias en los diferentes tipos de Violencia Contra la Mujer (VCM).***

  En esta tesis se han aplicado varios algoritmos de clasificación: Bosques aleatorios "Random Forest", Máquina de Soporte Vectorial "Support Vector Machines", Redes Bayesianas "Naive Bayes", Arból de decision "Decision Tree", "XGBoost Classifier" y Redes Neuronales. Nosotros decidimos aplicar estos algoritmos para clasificar noticias de alta dimensionalidad en los diferentes tipos de violencia contra la mujer; violencia física, violencia psicológica y violencia sexual. Sin embargo, no todos los algoritmos obtuvieron un buen rendimiento en su proceso de evaluación. Nosotros encontramos que en términos de eficiencia y desempeño el mejor clasificador de texto para el estudio de la VCM fue XGBoost. Sus altas medidas *"accuracy"* reflejan un buen grado de fiabilidad y confianza en la clasificación. Otro algoritmo que obtuvo un desempeño idóneo con poca diferencia de XGBoost fue "Random Forest". Por otra parte, los algoritmos que obtuvieron el más bajo desempeño fueron Redes Bayesianas y Árbol de Decisión sus valores de "accuracy" muestran que existe un bajo nivel de documentos clasificados correctamente evidenciando su bajo rendimiento al clasificar.

- ***Objetivo 4: Generar temas latentes a partir de noticias sobre la Violencia Contra la Mujer (VCM).***

  Dentro de la Minería de Texto podemos encontrar modelos probabilísticos que basados en un análisis bayesiano jerárquico permiten descubrir la estructura semántica subyacente de una colección de documentos. A este proceso se le conoce como Modelado de Temas y existen varios investigadores que han publicado muchos artículos aplicándolo en diferentes disciplinas. Una de la técnica más utilizada debido a su buen rendimiento es "Latent Dirichlet Allocation" (LDA) por lo que en esta investigación se presentó un enfoque aplicando LDA para generar un conjunto de temas latentes sobre la VCM. Se

logró descubrir relaciones entre datos y documentos de texto. Los temas obtenidos y los términos que los conforman revelaron de forma clara la naturaleza del tema. En consecuencia, se puede decir que LDA demostró un buen rendimiento en la extracción de estructuras semánticas de una larga colección de documentos convirtiéndolo en una técnica capaz de generar automáticamente temas representativos relacionados a la VCM.

- ***Objetivo 5: Extraer reglas de asociación que representen patrones frecuentes y así analizar las diferentes características asociadas a la Violencia Contra la Mujer (VCM).***

  Las Reglas de Asociación son una técnica muy popular aplicada para identificar patrones escondidos en un conjunto de datos. Entre los algoritmos más utilizados encontramos a Apriori un algoritmo que de forma automática permite extraer patrones de un conjunto de datos. Este ha sido aplicado en una variedad de investigaciones obteniendo buenos resultados. La VCM es un tema muy importante por lo que en esta investigación se decidió aplicar este algoritmo para determinar patrones sobre la VCM. A partir de noticias sobre casos reales de violencia se consiguió generar y reconocer asociaciones entre varias características de la VCM, entre ellas tenemos: quién es la víctima o el agresor más común, el tipo de arma más utilizada, entre otras. Estos resultados reflejan información relevante sobre episodios de violencia y podrían ser de gran utilidad para entender el gran alcance de esta problemática, el impacto negativo en las víctimas como en las familias sirviendo de apoyo en la creación o mejora de las leyes existentes para así contribuir a la prevención y reducción de la VCM.

5.2 TRABAJOS FUTUROS

A lo largo del desarrollo de esta tesis doctoral se ha tratado un problema de carácter social que se ha expandido a nivel mundial llamado Violencia Contra la Mujer. Después de concluir dicho estudio se ha observado que aún existen líneas de investigación que se pueden explorar. A continuación, se recogen las nuevas líneas de investigación a seguir:

- Se propone aplicar nuevos métodos de modelado de temas entre ellos “lda2vec”, además que se analice el rendimiento en comparación con el método LDA. De igual forma este proceso se podría llevar a cabo en texto en español como en texto en idioma inglés.

- Este estudio se enfocó al texto en idioma español, por lo que, se considera factible adaptar una metodología que permita aplicarlo a texto en idioma inglés. Así se podría evaluar el comportamiento sobre la VCM en países de habla inglesa. En este caso se deberían adaptar los diccionarios de violencia a los términos del nuevo idioma.

- Con respecto a la clasificación de texto se podría abrir una nueva vía para investigar otro tipo de violencia, no solo en la mujer, sino también estudiar la violencia que sufren los hombres, para lo cual sería necesario añadir o expandir los diccionarios de violencia existentes.

- Otra idea que surge es previa a la clasificación de documentos realizar un trabajo de clustering donde se agrupe los datos por tipo de violencia para luego examinar cuan eficiente son los resultados obtenidos en comparación de la aplicación de algoritmos de clasificación.

- Finalmente, en un trabajo futuro se podría aplicar reglas de asociación difusas y enfocarse en el estudio de otros tipos de violencia, u otros tipos de víctimas o características ya sea en textos extensos o cortos.

6 PUBLICACIÓN ASOCIADA A LA TESIS DOCTORAL

En este capítulo detallaremos la investigación científica publicada previo a la obtención de la titulación emitida por la Universidad de Granada.

Revista: International Journal of Data Science and Analytics

Url: https://doi.org/10.1007/s41060-023-00448-y

Título: Study of violence against women and its characteristics through the application of text mining techniques.

BIBLIOGRAFÍA

[1] C. García-moreno, "WHO Multi-country Study on Women's Health and Domestic Violence against Women."

[2] G. Krantz and C. Garcia-Moreno, "Violence against women," *J. Epidemiol. Community Health*, vol. 59, no. 10, pp. 818–821, 2005, doi: 10.1136/jech.2004.022756.

[3] U. Nations, *United Nations Report of the Fourth World Conference on Women*, no. September. 1995.

[4] L. American University, "The Elimination of Violence Against Women," *Al-Raida J.*, no. December, pp. 28–29, 1970, doi: 10.32380/alrj.v0i0.943.

[5] "WHO_violence against women a priority issue.pdf." .

[6] R. Espinoza Bonifaz, "Violencia contra la mujer. ¿Un problema de falta de normatividad penal o socio cultural?," *Vox juris*, vol. 37, no. 1, pp. 163–175, 2019, doi: 10.24265/voxjuris.2019.v37n1.12.

[7] J. D. Febro, Amor, "Explorando la ciberviolencia contra mujeres y niñas en Filipinas a través de Mining Online News Exploring cyber violence against women and girls in the Philippines," *Comunicar*, pp. 125–138, 2022.

[8] D. C. Garcia-Moreno, "Estimaciones mundiales y regionales de la violencia contra la mujer: prevalencia y efectos de la violencia conyugal y de la violencia sexual no conyugal en la salud," *Organ. Mund. la Salud*, p. 2, 2013, [Online]. Available: http://apps.who.int/iris/bitstream/10665/85243/1/WHO_RHR_HRP_13.06_spa.pdf.

[9] N. Gasman, L. Villa Torres, C. Moreno, and D. L. Billings, "Razones por las que no se denuncian los casos de abuso a la policía.," *Inf. Nac. sobre Violencia y Salud*, pp. 167–204, 2016, [Online]. Available: http://www.svri.org/nacional.pdf%5Cnhttp://www.paho.org/hq/index.php?option=com_docman&task=doc_download&gid=23947&Itemid=270.

[10] E. F. Arias, L. M. Vilcas Baldeón, and Y. Alberto Bueno, "Factores de riesgo de violencia a la mujer de parte del cónyuge," *Socialium*, vol. 3, no. 1, pp. 69–96, 2019, doi: 10.31876/sl.v3i1.67.

[11] B. Krahé, "Violence against women," *Curr. Opin. Psychol.*, vol. 19, pp. 6–10, 2018, doi: 10.1016/j.copsyc.2017.03.017.

[12] C. Hernández Ramos, V. Magro Servet, and J. P. Cuéllar Óton, "El maltrato psicológico. Causas, consecuencias y criterios jurisprudenciales. El problema probatorio," *Aequitas*, vol. 3, no. 7, pp. 27–53, 2014, [Online]. Available: https://www.scirp.org/html/33-1763258_99085.htm%0Ahttp://rua.ua.es/dspace/handle/10045/46929.

[13] H. Saldaña and G. de J. Gorjón, "Causas y consecuencias de la violencia familia," *Nuevo León". Justicia*, vol. 25, no. 38, pp. 189–214, 2020, [Online]. Available: http://revistas.unisimon.edu.co/index.php/justicia/article/view/4002/4935.

[14] E. Dytel *et al.*, “Universidad nacional hermilio valdizán escuela de posgrado,” 2019, [Online]. Available: http://repositorio.unheval.edu.pe/bitstream/handle/UNHEVAL/5319/PPE00220E88.pdf?sequence=1&isAllowed=y.

[15] L. Schrubbe, C. García-Moreno, L. Sardinha, and H. Stöckl, “Intimate partner violence against women during pregnancy: a systematic review and meta-analysis protocol for producing global and regional estimates,” *Syst. Rev.*, vol. 12, no. 1, 2023, doi: 10.1186/s13643-023-02232-2.

[16] R. Patró Hernández and R. M. Limiñana Gras, “Víctimas de violencia familiar: consecuencias psicológicas en hijos de mujeres maltratadas,” *An. Psicol.*, vol. 21, no. 1, pp. 11–17, 2005.

[17] United Nation-Women, “Intimate partner violence in five Caricom countries: Findings from national prevalence surveys on violence against women,” no. May, 2020, [Online]. Available: file:///C:/Users/inbalh/Downloads/20201009 CARICOM Research Brief 5.pdf.

[18] E. Persson, “Evaluating tools and techniques for web scraping,” *Degree Proj. Comput. Sci. Eng.*, vol. SECOND CYC, pp. 1–97, 2019, [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1415998/FULLTEXT01.pdf.

[19] V. Singrodia, A. Mitra, and S. Paul, “A Review on Web Scrapping and its Applications,” *2019 Int. Conf. Comput. Commun. Informatics, ICCCI 2019*, pp. 1–6, 2019, doi: 10.1109/ICCCI.2019.8821809.

[20] B. Zhao, “Web Scraping,” no. December, 2018, doi: 10.1007/978-3-319-32001-4.

[21] M. A. Khder, “Web scraping or web crawling: State of art, techniques, approaches and application,” *Int. J. Adv. Soft Comput. its Appl.*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[22] M. Khan, S. S. Khran, and Y. Alharbi, “Text Mining Challenges and Applications, A Comprehensive Review,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 20, no. 12, pp. 138–148, 2020, [Online]. Available: https://doi.org/10.22937/IJCSNS.2020.20.12.15.

[23] K. L.Sumathy and M. Chidambaram, “Text Mining: Concepts, Applications, Tools and Issues An Overview,” *Int. J. Comput. Appl.*, vol. 80, no. 4, pp. 29–32, 2013, doi: 10.5120/13851-1685.

[24] A. Alqahtani, “A Survey of Text Matching Techniques,” vol. 11, no. 1, pp. 6656–6661, 2021.

[25] G. Manias, A. Mavrogiorgou, A. Kiourtis, and D. Kyriazis, *SemAI: A Novel Approach for Achieving Enhanced Semantic Interoperability in Public Policies*, vol. 627. 2021.

[26] Q. Zu Yong Tang Vladimir Mladenovic, *Human Centered Computing 6th International Conference, HCC 2020 Virtual Event, December 14-15, 2020 Revised Selected Papers*. 2020.

[27] C. C. Aggarwal, *Mining text streams*, vol. 9781461432. 2012.

[28] S. Naithani and A. Kaushik, “A Comprehensive Study of Text Mining Approach.pdf A Comprehensive Study of Text Mining Approach,” *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 16, no. 2, p. 69, 2016.

[29] M. Khan, S. S. Khan, and Y. Alharbi, “Text Mining Challenges and Applications,

A Comprehensive Review," *Int. J. Comput. Sci. Netw. Secur.*, vol. 20, no. 12, pp. 138–148, 2020, doi: 10.22937/IJCSNS.2020.20.12.15.

[30] J. Jiang *et al.*, "We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %," *Intech*, vol. 34, no. 8, pp. 57–67, 2010, [Online]. Available: https://doi.org/10.1007/s12559-021-09926-6%0Ahttps://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics%0Ahttp://dx.doi.org/10.1016/j.compmedimag.2010.07.003.

[31] C. Rossi *et al.*, "Early detection and information extraction for weather-induced floods using social media streams," *Int. J. Disaster Risk Reduct.*, vol. 30, no. March, pp. 145–157, 2018, doi: 10.1016/j.ijdrr.2018.03.002.

[32] M. Afshar *et al.*, "Subtypes in patients with opioid misuse: A prognostic enrichment strategy using electronic health record data in hospitalized patients," *PLoS One*, vol. 14, no. 7, pp. 1–18, 2019, doi: 10.1371/journal.pone.0219717.

[33] H. Emami, H. Shirazi, and A. Abdollahzadeh Barforoush, "A semantic approach to cross-document person profiling in Web," *AI Commun.*, no. November, pp. 1–29, 2017, doi: 10.3233/aic-170472.

[34] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When Machine Learning Meets Privacy: A Survey and Outlook," *ACM Comput. Surv.*, vol. 54, no. 2, 2021, doi: 10.1145/3436755.

[35] D. Chandra Das, D. Saha, T. Kabir, P. Deb, and J. Bhowmik, "Analysis of Covid-19 Coverage in Bangladesh News Media Using Topic Modelling," *Int. J. Eng. Appl. Sci. Technol.*, vol. 5, no. 7, pp. 27–34, 2020, doi: 10.33564/ijeast.2020.v05i07.006.

[36] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, 2018, doi: 10.1109/MCI.2018.2840738.

[37] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 3730–3738, 2015, doi: 10.1109/ICCV.2015.425.

[38] S. Hu, Y. C. Liang, Z. Xiong, and D. Niyato, "Blockchain and Artificial Intelligence for Dynamic Resource Sharing in 6G and beyond," *IEEE Wirel. Commun.*, vol. 28, no. 4, pp. 145–151, 2021, doi: 10.1109/MWC.001.2000409.

[39] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J. Phys. Conf. Ser.*, vol. 1142, no. 1, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[40] J. Verbraeken, M. Wolting, and J. Katzy, "A Survey on Distributed Machine Learning," vol. 53, no. 2, 2020, doi: 10.1145/3377454.

[41] B. Mahesh, "Machine Learning Algorithms - A Review," no. January 2019, 2020, doi: 10.21275/ART20203995.

[42] A. Awaysheh, J. Wilcke, F. Elvinger, L. Rees, W. Fan, and K. L. Zimmerman, "Review of Medical Decision Support and Machine-Learning Methods," *Vet. Pathol.*, vol. 56, no. 4, pp. 512–525, 2019, doi: 10.1177/0300985819829524.

[43] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78,

2018, doi: 10.26438/ijcse/v6i10.7478.

[44] A. Patra and D. Singh, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms," *Int. J. Comput. Appl.*, vol. 75, no. 7, pp. 14–18, 2013, doi: 10.5120/13122-0472.

[45] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[46] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[47] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," *Proc. - 2018 10th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2018*, vol. 2, pp. 48–51, 2018, doi: 10.1109/IHMSC.2018.10117.

[48] F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, "Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors," *J. Mater. Chem. C*, vol. 3, pp. 10715–10722, 2015, doi: 10.1039/b000000x.

[49] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.

[50] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3797–3816, 2019, doi: 10.1007/s11042-018-6083-5.

[51] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[52] V. K. Vijayan, K. R. Bindu, and L. Parameswaran, "A comprehensive study of text classification algorithms," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 1109–1113, 2017, doi: 10.1109/ICACCI.2017.8125990.

[53] S. C. Dharmadhikari, Maya Ingle, and P. Kulkarni, "Empirical Studies On Machine Learning Based Text Classification Algorithms," *Adv. Comput. An Int. J.*, vol. 2, no. 6, pp. 161–169, 2011, doi: 10.5121/acij.2011.2615.

[54] W. B. Zulfikar, M. Irfan, C. N. Alam, and M. Indra, "The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, pp. 3–7, 2017, doi: 10.1109/CITSM.2017.8089231.

[55] E. D. Canedo and B. C. Mendes, "Software requirements classification using machine learning algorithms," *Entropy*, vol. 22, no. 9, 2020, doi: 10.3390/E22091057.

[56] T. B. Shahi, "Nepali News Classification using Naïve Bayes , Support Vector Machines and Neural Networks," pp. 18–22, 2018.

[57] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications,

challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[58] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[59] *Ensemble Machine Learning*. 2012.

[60] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.

[61] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-25.

[62] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1862–1874, 2015, doi: 10.1109/TPAMI.2014.2382106.

[63] M. Mahdianpari *et al.*, "Big Data for a Big Country: The First Generation of Canadian Wetland Inventory Map at a Spatial Resolution of 10-m Using Sentinel-1 and Sentinel-2 Data on the Google Earth Engine Cloud Computing Platform," *Can. J. Remote Sens.*, vol. 46, no. 1, pp. 15–33, 2020, doi: 10.1080/07038992.2019.1711366.

[64] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[65] A. Abid, W. Ali, M. S. Farooq, U. Farooq, N. S. Khan, and K. Abid, "Semi-automatic classification and duplicate detection from human loss news corpus," *IEEE Access*, vol. 8, pp. 97737–97747, 2020, doi: 10.1109/ACCESS.2020.2995789.

[66] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[67] P. Wang *et al.*, "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text," *IEEE Access*, vol. 8, pp. 97370–97382, 2020, doi: 10.1109/ACCESS.2020.2995905.

[68] J. J. Espinosa Zúñiga, "Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito," *Ing. Investig. y Tecnol.*, vol. 21, no. 3, pp. 1–16, 2020, doi: 10.22201/fi.25940732e.2020.21.3.022.

[69] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, no. August, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[70] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharm. Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, 2000, doi: 10.1016/S0731-7085(99)00272-1.

[71] A. P. Marugán, F. P. G. Márquez, J. M. P. Perez, and D. Ruiz-Hernández, "A survey of artificial neural network in wind energy systems," *Appl. Energy*, vol. 228, no. April, pp. 1822–1836, 2018, doi: 10.1016/j.apenergy.2018.07.084.

[72] M. Puri, A. Solanki, T. Padawer, S. M. Tipparaju, W. A. Moreno, and Y. Pathak, *Introduction to Artificial Neural Network (ANN) as a Predictive Tool for Drug Design, Discovery, Delivery, and Disposition: Basic Concepts and Modeling. Basic Concepts and Modeling*. Elsevier Inc., 2016.

[73] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, 2018, doi: 10.1016/j.heliyon.2018.e00938.

[74] B. Shah and B. H Trivedi, "Artificial Neural Network based Intrusion Detection System: A Survey," *Int. J. Comput. Appl.*, vol. 39, no. 6, pp. 13–18, 2012, doi: 10.5120/4823-7074.

[75] D. R. Nayak, "A Survey on Rainfall Prediction using Artificial Neural Network," no. July 2015, pp. 31–40, 2013, doi: 10.5120/12580-9217.

[76] S. Sah, "Machine Learning: A Review of Learning Types," *DOI 10.20944/preprints202007.0230.v1*, no. July, 2020, doi: 10.20944/preprints202007.0230.v1.

[77] T. Jiang, J. L. Gradus, and A. J. Rosellini, "ScienceDirect Supervised Machine Learning : A Brief Primer," *Behav. Ther.*, vol. 51, no. 5, pp. 675–687, 2020, doi: 10.1016/j.beth.2020.05.002.

[78] E. Cihan, E. Bostanci, and M. Serdar, "Machine Translation , Sentiment Analysis , Text Similarity , Topic Modelling , and Tweets : Understanding Social Media Usage Among Police and Gendarmerie Organizations," pp. 1–21, 2021.

[79] M. Hossain, A. Ahmed, and M. S. Rahman, "Topic Modelling : A Comparison of The Performance of Latent Dirichlet Allocation and LDA2vec Model on Bangla Newspaper," no. September, pp. 27–28, 2019.

[80] P. V. Poojitha and R. R. K. Menon, "Document Representations to Improve Topic Modelling," pp. 18–25, 2020.

[81] C. Jacobi, W. Van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digit. Journal.*, vol. 4, no. 1, pp. 89–106, 2016, doi: 10.1080/21670811.2015.1093271.

[82] Q. Zhou and C. Zhang, "Emotion evolutions of sub-topics about popular events on microblogs," *Electron. Libr.*, vol. 35, no. 4, pp. 770–782, 2017, doi: 10.1108/EL-09-2016-0184.

[83] E. S. Negara, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," pp. 386–390, 2019.

[84] C. Sharma, S. Sharma, and Sakshi, "Latent DIRICHLET allocation (LDA) based information modelling on BLOCKCHAIN technology: a review of trends and research patterns used in integration," *Multimed. Tools Appl.*, vol. 81, no. 25, pp. 36805–36831, 2022, doi: 10.1007/s11042-022-13500-z.

[85] A. Telikani, A. H. Gandomi, and A. Shahbahrami, "A survey of evolutionary computation for association rule mining," *Inf. Sci. (Ny).*, vol. 524, pp. 318–352, 2020, doi: 10.1016/j.ins.2020.02.073.

[86] S. K. Solanki and J. T. Patel, "A survey on association rule mining," *Int. Conf.*

Adv. Comput. Commun. Technol. ACCT, vol. 2015-April, pp. 212–216, 2015, doi: 10.1109/ACCT.2015.69.

[87] J. K. K. Chung, "Mining from Health Big Data," *Wirel. Pers. Commun.*, vol. 105, no. 2, pp. 691–707, 2019, doi: 10.1007/s11277-018-5722-5.

[88] A. Toumi, N. Gribaa, and W. B. A. Karaa, "Mining biomedical texts based on statistical method and association rules," *2021 Int. Conf. Women Data Sci. Taif Univ. WiDSTaif 2021*, 2021, doi: 10.1109/WIDSTAIF52235.2021.9430200.

[89] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Commun. Surv. Tutorials*, vol. 15, no. 3, pp. 1136–1159, 2013, doi: 10.1109/SURV.2012.100412.00017.

[90] P. K. Singh, A. K. Kar, Y. Singh, M. H. Kolekar, and S. Tanwar, *of ICRIC*. 2019.

[91] P. P. Shinde, "A Review of Machine Learning and Deep Learning Applications," *2018 Fourth Int. Conf. Comput. Commun. Control Autom.*, pp. 1–6, 2018.

[92] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision : A Brief Review," vol. 2018, 2018.

[93] X. Yin, "Ethical Dilemma of Using Natural Language Processing (NPL) Machine Learning Method in Auditing Company Internal Communication," vol. 7, no. 3, pp. 16–20, 2023, doi: 10.25236/IJNDES.2023.070303.

[94] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, "ScienceDirect A Review Review on on recent recent research research in in information information retrieval retrieval," *Procedia Comput. Sci.*, vol. 201, pp. 777–782, 2022, doi: 10.1016/j.procs.2022.03.106.

[95] A. Tahseen Ali, H. S. Abdullah, and M. N. Fadhil, "Voice recognition system using machine learning techniques," *Mater. Today Proc.*, no. xxxx, 2022, doi: 10.1016/j.matpr.2021.04.075.

[96] S. Das, A. Dey, A. Pal, and N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *Int. J. Comput. Appl.*, vol. 115, no. 9, pp. 31–41, 2015, doi: 10.5120/20182-2402.

[97] A. Handa, A. Sharma, and S. K. Shukla, "Machine learning in cybersecurity: A review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, pp. 1–7, 2019, doi: 10.1002/widm.1306.

[98] M. Nobakht, V. Sivaraman, and R. Boreli, "A host-based intrusion detection and mitigation framework for smart home IoT using OpenFlow," *Proc. - 2016 11th Int. Conf. Availability, Reliab. Secur. ARES 2016*, pp. 147–156, 2016, doi: 10.1109/ARES.2016.64.

[99] A. Finogeev, A. Finogeev, L. Fionova, A. Lyapin, and K. A. Lychagin, "Intelligent monitoring system for smart road environment," *J. Ind. Inf. Integr.*, vol. 15, no. April 2018, pp. 15–20, 2019, doi: 10.1016/j.jii.2019.05.003.

[100] J. Li, Q. Huang, S. Ren, L. Jiang, B. Deng, and Y. Qin, "A novel medical text classification model with Kalman filter for clinical decision making," *Biomed. Signal Process. Control*, vol. 82, no. September 2022, p. 104503, 2023, doi: 10.1016/j.bspc.2022.104503.

[101] J. Sharma, K. Sharma, K. Garg, and A. K. Sharma, "Product recommendation system a comprehensive review," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012021.

[102] S. Giovany, A. Putra, A. S. Hariawan, and L. A. Wulandhari, “Machine Learning and SIFT Approach for Indonesian Food Image Recognition,” *Procedia Comput. Sci.*, vol. 116, pp. 612–620, 2017, doi: 10.1016/j.procs.2017.10.020.

[103] V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. D. Ramkteke, “Machine learning in agriculture domain: A state-of-art survey,” *Artif. Intell. Life Sci.*, vol. 1, no. September, p. 100010, 2021, doi: 10.1016/j.ailsci.2021.100010.

[104] F. N. Patel and N. R. Soni, “Text mining: A Brief survey,” *Int. J. Adv. Comput. Res.*, vol. 2, no. 6, pp. 243–248, 2012, [Online]. Available: http://www.theaccents.org/ijacr/papers/conference/icett2012/43.pdf.

[105] M. Allahyari *et al.*, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” *arXiv*, 2017.

[106] V. Kumar and S. Minz, “Feature Selection : A literature Review,” vol. 4, no. 3, 2014, doi: 10.6029/smartcr.2014.03.007.

[107] D. E. C. Na and C. Hipertensiva, “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title.”

[108] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, “Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods,” *Appl. Soft Comput. J.*, vol. 86, p. 105836, 2020, doi: 10.1016/j.asoc.2019.105836.

[109] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, “SMOTE for Handling Imbalanced Data Problem : A Review,” *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, pp. 1–8, 2021, doi: 10.1109/ICIC54025.2021.9632912.

[110] W. Pei, B. Xue, M. Zhang, L. Shang, X. Yao, and Q. Zhang, “A Survey on Unbalanced Classification: How Can Evolutionary Computation Help?,” *IEEE Trans. Evol. Comput.*, vol. PP, p. 1, 2023, doi: 10.1109/TEVC.2023.3257230.

[111] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of Classification Methods on Unbalanced Data Sets,” *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.

[112] S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[113] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “snopes.com: Two-Striped Telamonia Spider,” *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002, [Online]. Available: https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

[114] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost: A hybrid approach to alleviating class imbalance,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 40, no. 1, pp. 185–197, 2010, doi: 10.1109/TSMCA.2009.2029559.

[115] M. H. A. Hamid, M. Yusoff, and A. Mohamed, “Survey on Highly Imbalanced Multi-class Data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 211–229, 2022, doi: 10.14569/IJACSA.2022.0130627.

[116] V. Gupta and G. S. Lehal, “A Survey of Text Mining Techniques and Applications

- Volume 1, No. 1, August 2009 - JETWI," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, 2009, [Online]. Available: http://www.jetwi.us/index.php?m=content&c=index&a=show&catid=165&id=969.

[117] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "ECG beat classification using machine learning techniques," *Int. J. Biomed. Eng. Technol.*, vol. 26, no. 1, pp. 32–53, 2018, doi: 10.1504/IJBET.2018.089255.

[118] C. Zhang and Y. Lu, "Study on artificial intelligence: The state of the art and future prospects," *J. Ind. Inf. Integr.*, vol. 23, no. March, p. 100224, 2021, doi: 10.1016/j.jii.2021.100224.

[119] E. M. A. Stephanie, L. G. B. Ruiz, M. A. Vila, and M. C. Pegalajar, "Study of violence against women and its characteristics through the application of text mining techniques," *Int. J. Data Sci. Anal.*, 2023, doi: 10.1007/s41060-023-00448-y.

[120] H. Alsaif and T. Alotaibi, "Arabic Text Classification using Feature-Reduction Techniques for Detecting Violence on Social Media," vol. 10, no. 4, pp. 77–87, 2019.

[121] P. Barberá, A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler, "Automated Text Classification of News Articles: A Practical Guide," *Polit. Anal.*, vol. 29, no. 1, pp. 19–42, 2021, doi: 10.1017/pan.2020.8.

[122] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6595–6604, 2022, doi: 10.1016/j.jksuci.2022.03.020.

[123] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A Deep Learning Approach to Network Intrusion Detection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 1, pp. 41–50, 2018, doi: 10.1109/TETCI.2017.2772792.

[124] L. Neubauer, I. Straw, E. Mariconti, and L. M. Tanczer, "A Systematic Literature Review of the Use of Computational Text Analysis Methods in Intimate Partner Violence Research," *J. Fam. Violence*, no. 0123456789, 2023, doi: 10.1007/s10896-023-00517-7.

[125] M. Hofer, G. Strauß, K. Koulechov, and A. Dietz, "Definition of accuracy and precision-evaluating CAS-systems," *Int. Congr. Ser.*, vol. 1281, pp. 548–552, 2005, doi: 10.1016/j.ics.2005.03.290.

[126] A. Menditto, M. Patriarca, and B. Magnusson, "Understanding the meaning of accuracy, trueness and precision," *Accredit. Qual. Assur.*, vol. 12, no. 1, pp. 45–47, 2007, doi: 10.1007/s00769-006-0191-z.

[127] P. Hajibabaee *et al.*, "Offensive Language Detection on Social Media Based on Text Classification," *2022 IEEE 12th Annu. Comput. Commun. Work. Conf. CCWC 2022*, pp. 92–98, 2022, doi: 10.1109/CCWC54503.2022.9720804.

[128] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 54, no. 3, 2020, [Online]. Available: http://arxiv.org/abs/2004.03705.

[129] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, pp. 221–232, 2017, doi: 10.22364/bjmc.2017.5.2.05.

[130] D. Campos, R. R. Silva, and J. Bernardino, “Text mining in hotel reviews: Impact of words restriction in text classification,” *IC3K 2019 - Proc. 11th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, no. January, pp. 442–449, 2019, doi: 10.5220/0008346904420449.

[131] L. Li, T. T. Goh, and D. Jin, “How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis,” *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4387–4415, 2020, doi: 10.1007/s00521-018-3865-7.

[132] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, “Comparing automated text classi fi cation methods,” vol. 36, pp. 20–38, 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[133] Y. Xiang *et al.*, “Artificial Intelligence-Based Diagnosis of Diabetes Mellitus: Combining Fundus Photography with Traditional Chinese Medicine Diagnostic Methodology,” *Biomed Res. Int.*, vol. 2021, 2021, doi: 10.1155/2021/5556057.

[134] C. Zong, R. Xia, and J. Zhang, “Topic Model,” *Text Data Min.*, pp. 145–162, 2021, doi: 10.1007/978-981-16-0100-2_7.

[135] O. Lomoschitz, “María Ibáñez de Opacua Lomoschitz,” 2019.

[136] M. Maryamah, A. Z. Aritin, R. Sarno, and R. W. Sholikah, “Enhanced Topic Modelling using Dictionary For,” pp. 219–223, 2019.

[137] M. B. Mutanga and A. Abayomi, “Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach,” *African J. Sci. Technol. Innov. Dev.*, vol. 0, no. 0, pp. 1–10, 2020, doi: 10.1080/20421338.2020.1817262.

[138] G. Brookes and T. McEnery, “The utility of topic modelling for discourse studies: A critical evaluation,” *Discourse Stud.*, vol. 21, no. 1, pp. 3–21, 2019, doi: 10.1177/1461445618814032.

[139] X. Cheng *et al.*, “Topic modelling of ecology, environment and poverty nexus: An integrated framework,” *Agric. Ecosyst. Environ.*, vol. 267, no. March, pp. 1–14, 2018, doi: 10.1016/j.agee.2018.07.022.

[140] A. Moodley and V. Marivate, “Topic Modelling of News Articles for Two Consecutive Elections i n South Africa,” pp. 131–136, 2019.

[141] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association in Large Databases,” *Proc. 1993 ACM SIGMOD Int. Conf. Manag. data - SIGMOD ’93*, pp. 207–216, 1993.

[142] M. H. Tseng and H. C. Wu, “Investigating health equity and healthcare needs among immigrant women using the association rule mining method,” *Healthc.*, vol. 9, no. 2, 2021, doi: 10.3390/healthcare9020195.

[143] R. Xu and F. Luo, “Risk prediction and early warning for air traffic controllers’ unsafe acts using association rule mining and random forest,” *Saf. Sci.*, vol. 135, no. August 2020, p. 105125, 2021, doi: 10.1016/j.ssci.2020.105125.

[144] V. Jo, *Introduction*, vol. 36, no. 2. 2019.

[145] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, “Non-query-based pattern mining and sentiment analysis for massive microblogging online texts,” *IEEE Access*, vol. 8, pp. 78166–78182, 2020, doi: 10.1109/ACCESS.2020.2990461.

[146] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, "Preprocessing Techniques for Text Mining -An Overview," vol. 5, no. 1, pp. 7–16.

[147] J. Sanmartín, "¿Qué es la violencia? Una aproximación al concepto y a la clasificación de la violencia," *Daimon Rev. Filos.*, vol. 0, no. 42, pp. 9–22, 2007.

[148] Y. Rodríguez López, B. Arenia Aguiar Gigato, I. Garcia Alvarez, and sldcu Lic Yahira Rodríguez, "CONSECUENCIAS PSICOLÓGICAS DEL ABUSO SEXUAL INFANTIL PSYCHOLOGICAL CONSEQUENCES OF INFANT SEXUAL ABUSE Correspondencia puede ser remitida a: psicoart@infomed," *Asunción (Paraguay)*, vol. 9, no. 1, pp. 58–68, 2012.

[149] N. S. Purba and R. Nooraeni, "Using LDA for Innovation Topic of Technology : Quantum Dots Patent Analysis," no. January, 2020, doi: 10.4108/eai.2-8-2019.2290336.

[150] S. Mohammad, J. Jalali, and H. W. Park, "OUP accepted manuscript," *Digit. Scholarsh. Humanit.*, 2020, doi: 10.1093/llc/fqaa012.