Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Identificación automática de eventos delictivos en noticias periodísticas

Tesis presentada para obtener el título de:
Maestría en Bases de Datos y Recuperación de Información

Presenta:
Yadira Laureano de Jesús

Asesor de Tesis:
Dr. Guillermo De Ita Luna

Co-asesora de Tesis:
Dra. Mireya Tovar Vidal

Junio 2020

Dedicatoria

A mi hijo Héctor Alfredo, quien motiva e impulsa mi vida día a día, a mis padres Elías y Josefina quienes me apoyan incondicionalmente, por el cariño que me brindan y por sus consejos que han sabido guiarme para ser una mejor persona.

Agradecimientos

A mi familia por el apoyo que me brindaron en cada momento.

A mis asesores de tesis quienes me apoyaron constantemente en su elaboración.

A mis amigos por apoyarme e inspirarme para culminar esta etapa de mi vida.

A Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico para poder obtener este grado académico.

A todos aquellos que de alguna forma me apoyaron e inspiraron en este proyecto.

Resumen

Existen varios tipos de eventos delictivos, los cuales ponen en riesgo a los individuos en espacios públicos, en la comunidad o en algún otro espacio. Es por ello que es importante saber acerca de estos eventos y principalmente para las autoridades puesto que estas imparten la ley, siendo de gran importancia para que tomen medidas de seguridad y apliquen las políticas de prevención de estos eventos delictivos. Por lo cual este trabajo de investigación tiene como objetivo el desarrollo de un modelo basado en redes neuronales con aprendizaje profundo para la clasificación automática de noticias de eventos delictivos (homicidio, secuestro, asalto, suicidio y violación), las cuáles serán extraídas de la página local de noticias: Milenio, estas noticias serán utilizadas para realizar las pruebas y para el entrenamiento se utiliza un corpus de encabezados de noticias periodísticas de Twitter. En la evaluación de la clasificación se obtuvo una exactitud global de clasificación del 77 %, en precisión un 73 %, en exhaustividad se obtuvo un 62 % y en la métrica F_1 se obtuvo un 77 %. Posteriormente se realiza un análisis en las noticias, en este análisis se obtiene una visualización de las palabras más representativas a cada evento delictivo, además de la obtención de las zonas de alto riesgo en la república mexicana y finalmente se muestran los datos que representan a las víctimas y culpables en cada uno de estos eventos delictivos, así como el sexo y la edad.

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Capítulo 1

Introducción

A través de los años el uso de la tecnología ha aumentado considerablemente y con ello las personas acceden a la información de manera diferente, por ejemplo: las personas más jóvenes prefieren acceder a noticias a través de internet (Facebook, Twitter, etc.); mediante el uso de celulares, computadoras o tabletas. Mientras que las personas adultas prefieren usar otros medios; como por ejemplo: la televisión, prensa escrita y radio. Todo esto lo hacen para estar informados de lo que pasa en el mundo. Es por ello que las noticias son una fuente esencial de información para dar a conocer los hechos más trascendentes que ocurren u ocurrieron durante un periodo y ubicación especifica. La idea principal de las noticias, es que al lector le agrade lo que lee y sobre todo que la información que se proporcione sea de su interés. Las noticias se basan en los temas de la realidad actual, algunos de los temas son: política, deporte, policíacas, economía, entre otras. Uno de los temas más frecuentes dentro de las noticias, es el de policíacas, estas noticias giran en entorno a investigaciones de un delito convirtiéndose en el foco de atención de los lectores dejando a un lado los temas como política y economía. Las noticas con el tema del delito, genera miedo y preocupación al lector puesto que ponen en peligro su seguridad y la seguridad de sus familiares. Por lo tanto es de suma importancia analizar las noticias, dado que al hacer un análisis de las noticias se pueden obtener los eventos delictivos con más incidencia de ocurrencia por país, estado o entidad, al igual que el mes, año, etc. Además de que se pueden obtener las zonas con mayor índice de peligro y sobre todo el delito con más ocurrencia. Al obtener un análisis reforzado se puede alertar a las

comunidades más vulnerables y así tomar estrictas medidas de seguridad por parte de las autoridades. En este trabajo de investigación se pretende hacer un análisis de noticias para categorizar por evento delictivo en los estados de México, además de obtener datos que puedan ser de gran ayuda para la seguridad de las personas y sobre todo para la prevención del delito.

1.1. Planteamiento del problema

El análisis de las noticias periodísticas es esencial para la humanidad, dado que con ellas podemos saber lo que ha pasado y lo que está pasando en cualquier parte del mundo, además de que nos proporcionan información muy impórtate, para cierto país, estado o comunidad y de temas en específico tales como: política, economía, eventos delictivos, etc. Los delitos criminales cada vez están más presentes en la actualidad, cada año aumentan significativamente, con ello la inseguridad y el miedo en las personas. Es por ello que este trabajo de investigación se centra en la clasificación de noticias de los eventos delictivos, tales como: homicidio, secuestro, asalto, suicidio y violación. Para tal clasificación se utilizará un corpus de noticias de Twitter, además del análisis de las noticias periodistas obtenidas de la página local: Milenio, en donde se analizará la noticia para así obtener los elementos principales de esta, por ejemplo: Lugar, sexo y edad del suceso, centrándonos en los temas de eventos delictivos. Al obtener una clasificación de estos eventos y los elementos principales que componen a cada noticia, se podrá evaluar las zonas riesgo, además del delito presente en cada estado, así como el sexo y edad de culpables y víctimas. Con el fin de alertar a la comunidad para que tome medidas de seguridad y principalmente a las autoridades.

1.2. Objetivos

Nuestros objetivos planteados, se dividen en dos partes, el objetivo general que define el alcance de la investigacón y los objetivos específicos que determinan las tareas a realizar.

1.2.1. Objetivo General

Seleccionar y aplicar un algoritmo basado en redes neuronales para la clasificación automática de noticias de eventos delictivos y la detección automática de zonas de riesgo.

1.2.2. Objetivos Específicos

1. Crear un corpus anotado de noticias de eventos delictivos.
2. Aplicar un algoritmo basado en redes neuronales para la clasificación de eventos delictivos.
3. Aplicar un algoritmo de reconocimiento de entidades con nombre para detectar zonas de riesgo.
4. Evaluar los resultados obtenidos.

1.3. Antecedentes

En los últimos años el acceso a la información ha cambiado como se menciona en Yuste [38]. Los medios sociales (redes sociales, blogs, etc.) están presentes en la vida diaria de los jóvenes. Los cuales son, sin duda, su espacio natural, desde que realizan actividades diversas, así como establecer conversaciones con su comunidad más cercana e informarse de las noticias. Se informan a través: de los perfiles de los medios digitales los cuales se han abierto a estos espacios de información, precisamente para llegar a estos públicos mucho más reticentes por un medio convencional, además en medida que aumenta la edad en las personas, aumenta su interés por las noticias. El Internet es el medio preferido por los adultos más jóvenes a la hora de consultar la actualidad informativa. El 43 % elige esta fuente para mantenerse informado de las últimas noticias, mientras que un 35 % prefiere la televisión. Si se comparan los medios a nivel global, la televisión mantiene su estatus como medio de comunicación (elegido por el 52 %), si bien Internet le sigue a la zaga (49 %). Muy por detrás de ambos, lo que demuestra su falta de "adeptos", son la radio (27 %) y la prensa escrita (14 %).

El Instituto Nacional de Estadística y Geografía (INEGI) desarrolló desde 1996 el catalogo único de delitos mencionados en SSP y INEGI [30]. El INEGI presenta la Clasificación Estadística de Crímenes (CED), cada uno de los niveles de agrupación cuenta con un código o clave que permite recolectar, integrar, procesar y presentar resultados estadísticos para cada nivel jerárquico o de agrupación. Se planteó la división estructural del delito en cinco niveles de clasificación estadística: grupo principal, subgrupo, grupo unitario clase de delito y delito. Partiendo de lo general a lo particular, a efecto de facilitar la clasificación e identificación de cada uno de los delitos, con la finalidad de concretar el principio de flexibilidad y versatilidad expresado respecto de los usos de esta clasificación, y de las necesidades que busca satisfacer. La clasificación del delito está compuesta por tres grupos principales:

1-2 Delitos contra las personas.

3-4 Delitos contra la sociedad.

5-6 Delitos contra el Estado.

Para cada uno de estos grupos principales se distinguen los cuatro niveles jerárquicos mencionados anteriormente.

La prevención del delito Koloffon et al. [13] es un componente central en toda estrategia de control del crimen y la violencia. Explicar su relevancia es sencillo: las políticas de prevención atienden factores de riesgo presentes en el individuo, en la familia, en los espacios públicos, en la comunidad, para evitar que se traduzcan en actos criminales. En pocas palabras, implica actuar antes de que se infrinja la ley y se lastime a una persona o grupo de ellas. La prevención exitosa, sin embargo, es una empresa compleja que requiere que distintos prerrequisitos estén satisfechos para que las políticas e intervenciones públicas puedan tocar esos puntos de riesgo con eficacia

Se consideran cinco eventos delictivos de SSP y INEGI [30] los cuales son: homicidio, secuestro, asalto, suicidio y violación. El reconocimiento automático de estos cinco eventos delictivos, extraídos de las noticias diarias en español por medio de periódicos locales online, es crucial para que el gobierno tome decisiones sobre la implementación de políticas y estrategias de prevención del delito para evitar eventos violentos en el futuro cercano.

En Reyes-Ortiz y Bravo [24] presentan un enfoque para mejorar los patrones

con información morfológica y categorías POS para reconocer y extraer eventos criminales de noticias publicadas en periódicos digitales mexicanos. Se consideran seis eventos criminales, de los grupos principales del CED (Clasificación estadística de los delitos) las cuales son: homicidio, violación, asalto, suicidio, secuestro y explotación sexual. Los patrones de las noticias en español utilizan un método semisupervisado, se mejoran con información lingüística (categorías morfológicas y POS) para reconocer eventos criminales de noticias en español no etiquetadas. El corpus de capacitación está compuesto por 1600 noticias en español para cada evento criminal, por lo tanto, se cuenta con 9600 titulares de noticias criminales en el corpus de capacitación. Se utilizaron las reglas de JAPE para caracterizar categorías de eventos criminales con el fin de anotarlas en el texto. Finalmente se realizó una evaluación utilizando una medida de *Macro* F_1 que considera la precisión y *recall* de eventos de extracción.

En este trabajo de tesis se propone el desarrollo de un modelo de clasificación automática de eventos delictivos periodísticos, el cual se basará en la clasificación por medio de redes neuronales profundas y la detección de zonas de riesgo.

1.4. Justificación

En la revisión de la literatura científica, explicada en el CAPÍTULO II, enfocados en los trabajos relacionados a los objetivos de este tema de investigación, no se ha encontrado investigaciones que permitan realizar la clasificación de eventos delictivos en noticias periodísticas a través de redes neuronales, con la clasificación de los eventos sobre: homicidio, secuestro, asalto, suicidio y violación, en noticias periodísticas mexicanas. Además de que no se han encontrado estudios relacionados con el aprendizaje profundo y redes neuronales para la clasificación de estos eventos delictivos en nuestro país, tampoco en el análisis de la información basándose en los resultados que se obtienen. El uso de estas técnicas para el análisis de información, permite de manera oportuna la identificación de los elementos principales de las noticias, así como patrones en un conjunto de datos para la clasificación de estos eventos delictivos.

1.5. Metodología

El desarrollo de este trabajo, se basara en la ciencia de datos, la cual nos proporciona los pasos que se deben seguir para hacer una implementación más fácil, más precisa o acertada. Uno de los puntos más importantes de estos pasos es la iniciación del proyecto, donde se define el proyecto a abordar y se finaliza con la ejecución de acciones.

Para la realización de este trabajo se realizó un análisis exhaustivo de trabajos anteriores, los cuales se basaron en los temas de interés, para tener un punto de referencia y así poder inicializar el trabajo. A fin de cumplir con el objetivo del trabajo, se empezó con la obtención de un corpus de noticias de la página local: Milenio, en seguida se aplica la red neuronal. Para el entrenamiento de esta red se utiliza un corpus de encabezados de noticias de la red social: Twitter y posteriormente se prueba la red neuronal con el corpus obtenido de la página local Milenio, finalmente se evalúa el modelo para tener una certeza de que tan buenos son los resultados. Al tener una buena clasificación, se procede a obtener los elementos principales de las noticias y con ello realizar un análisis que nos proporcione información importante de cada evento delictivo.

1.6. Distribución del trabajo de tesis

El presente trabajo de tesis se organiza de la siguiente manera:

- Capítulo 1 Introducción. Se muestra el planteamiento del problema, objetivos, antecedentes, justificación seguida de la metodología.
- Capítulo 2 Estado del Arte. Se describen las contribuciones realizadas por diversos autores en las tareas de “clasificación” y “redes neuronales” puntos clave de desarrollo de este trabajo de tesis.
- Capítulo 3 Marco teórico. Se exponen las bases teóricas de la investigación, útil para comprender el significado del contenido expuesto.
- Capítulo 4 Diseño. Se muestran los modelos propuestos para la resolución del problema expuesto.

- Capítulo 5 Resultados. Se exponen a detalle los resultados obtenidos por los modelos propuestos.
- Capítulo 6 Conclusiones. En esta sección son mostradas las conclusiones obtenidas al realizar este trabajo de tesis y exponemos el trabajo a futuro.

Capítulo 2

Estado del arte

En este capítulo, se presenta un estudio detallado del estado del arte con el objetivo de conocer investigaciones anteriores en las tareas de clasificación de noticias periodísticas, las cuales se describen a continuación.

En Valero [32] presenta una investigación la cual consiste en extraer información sobre desastres naturales a partir de noticias en español, las cuales fueron obtenidas de varios periódicos mexicanos que están disponibles en Internet (publicados entre los años de 1996 a 2004), de donde se limitó el dominio de extracción a sólo cinco clases de eventos: Forestal, huracán, inundación, sequía y sismo. Por lo cual se propone recurrir a un análisis mediante expresiones regulares para detectar las entidades. Se plantea una arquitectura donde se define un conjunto de componentes que en una primera etapa identifican fragmentos del texto con posibilidad de ser extraídos, y posteriormente deciden cuáles de estos fragmentos son relevantes al dominio. La característica principal de la arquitectura es el escaso uso de recursos lingüísticos, los cuales son reemplazados por métodos de aprendizaje supervisado (i.e, Naïve Bayes, C4.5, k-vecinos más cercanos y máquinas de soporte vectorial) los cuales fueron evaluados con la validación cruzada con 10 pliegues. Finalmente se presenta una aplicación real; esta aplicación permitió demostrar lo útil de la arquitectura, esto por comparar los resultados del sistema contra un trabajo de extracción realizado de forma manual, donde se comprobó que la propuesta puede ser útil para incrementar el conocimiento del caso de estudio.

En Lucía y Alvarado [14] presenta un trabajo donde se usa Twitter como fuente

de datos proporcionados por la facultad de Geología de la Universidad Complutense de Madrid y se propone desarrollar un método para analizar el texto de un conjunto de tweets. El método permite clasificar dichos tweets en las siguientes clases: tráfico, contaminación, o tráfico y contaminación o ninguna de estas. Empleando varios algoritmos de clasificación supervisada, que fueron previamente entrenados. Se estudiaron los siguientes cuatro algoritmos, Naïve Bayes multiclase, arboles de decisión, k-vecinos más cercanos y máquina de soporte vectorial, para obtener la exactitud de cada uno, y analizar cuál es el mejor algoritmo de clasificación para este caso de estudio. Las principales herramientas utilizadas son mongo y Python, que se empleó principalmente para los datos preprocesamiento y elaboración de minería de texto. Realizando la evaluación con las medidas de exactitud, recall, precisión, y F-score. En los resultados obtenidos, con el algoritmo de máquina de soporte vectorial, se logra un valor de exactitud de 85.22 % para la clasificación de eventos de tráfico y no tráfico. Además, se realizó la clasificación multiclase, donde se obtuvo un valor de exactitud de 78.84 %.

En la memoria presentada por Silva [28] profundiza sobre el nivel de criminalidad presente en las noticias policiales las cuales son recopiladas durante seis meses, desde julio a diciembre del año 2011, las cuales contiene 23,726 noticias. Para ello se trabaja en el diseño y la aplicación de una metodología (CRISP-DM) que permita de forma eficiente y estandarizada realizar las distintas tareas que involucra el análisis de noticias. Se plantea el diseño de un mecanismo metodológico para el procesamiento de los datos contenidos en grandes colecciones de noticias (textos) y la extracción de conocimiento útil, con el objetivo de utilizar dicho conocimiento para un posterior análisis de los niveles de cobertura policial y su relación con variables de fuentes externas como las estadísticas de casos de delitos reales. Se distinguieron siete temáticas las cuales son: delitos sexuales, incendios, drogas, disturbios, homicidios, tránsito y robos, las cuales presentan diferentes niveles de cobertura entre sí, así como también según la región y según el medio de prensa. Se estudiaron dos modelos de clasificación distintos: Naïve Bayes y K-nn, seleccionando el modelo con mejor desempeño basado en el método de evaluación cruzada k-fold (con k=10) y se comparan las medidas de recall, precisión, F-measure y accuracy. Finalmente se construye un prototipo de una herramienta de visualización de datos georreferenciados

utilizando como base el API de JavaScript de Google Maps.

En Álvaro [15] se presenta un trabajo donde se compararon algunas técnicas de clasificación y preprocesamiento de textos, centrando la mayor parte en un problema específico de clasificación conocido como análisis de sentimientos. Para ello, previamente se estudiaron las metodologías existentes de clasificación tales como arboles de decisión, Random Forest, Máxima Entropía, SVC Kernel Linear, NB Multinomial y Naïve Bayes; con el fin de comprender las limitaciones y ventajas de cada uno. Después, se escogieron algunas de estas metodologías y se realizarán pruebas sobre dos bases de datos alojadas en la web, la primera se llama Sentiment Labelled Sentences Data Set. Ésta base de datos está formada por tres documentos de diferentes contextos (Yelp, Amazon e IMDB). Cada documento contiene mil textos clasificados en inglés, de los cuales la mitad son positivos y la otra mitad son negativos, Movie Review Data Set, la cual está formada por 50,000 textos clasificados, la mitad como positivos y la otra mitad como negativos y otros 50,000 textos no clasificados y la base de datos de incidencias, proporcionada por la empresa Cognodata Consulting, consta de 57,735 registros. Tratando de comparar las diferentes metodologías e intentando finalmente construir el clasificador que obtenga mejores resultados. Se realizó la evaluación de precisión y rendimiento.

El siguiente apartado se muestra algunos estudios relacionados con *Machine Learning* y *Deep Learning* (aprendizaje profundo).

En Rama [22] presenta su proyecto el cual consiste en la aplicación de técnicas de Machine Learning a la seguridad, se pretende demostrar la utilidad de Machine Learning y su aplicación a la seguridad informática, utilizando algoritmos de aprendizaje supervisado, concretamente usando algoritmos de clasificación (Random Tree, Random Forest, J48, regresión logística, SVM y Naïve Bayes), realizando la evaluación con precisión. En donde se utiliza Weka y en Python el desarrollo de un script, el cual permite realizar el tratamiento del conjunto de datos y la posterior utilización de un modelo predictivo para la detección de conexiones maliciosas. El conjunto de datos utilizado es “KDD Cup 1999” que contiene un conjunto de conexiones clasificadas como normales o como un ataque determinado.

En el artículo presentado por Montañés et al. [16] describen la participación de ITAINNOVA en la tarea de análisis de sentimiento a nivel de Tweet dentro del taller

TASS 2018. El trabajo pretende explorar modelos presentes en el estado del arte actual del aprendizaje profundo aplicado al modelado y clasificación de texto. Se ha analizado el uso de modelos de redes convolucionales (CNN), Long short Term Memory (LSTM), LSTM bidireccionales (BI-LSTM) y una aproximación híbrida entre CNN y LSTM para su uso en el análisis de sentimiento en Twitter, basado exclusivamente en la variedad de español hablado en España, utilizando para ello el conjunto de datos: InterTASS ES junto con un subconjunto del corpus general utilizado desde las primeras ediciones, que permitirá predecir la polaridad de los tweets en base a cuatro niveles: P (Positiva), N (Negativa), NEU (Neutra), NONE (sin opinión). Ellos optaron por la combinación CNNLSTM ya que integra los beneficios de ambos modelos. Finalmente en su trabajo presentan los resultados obtenidos de la evaluación con *Macro-*F_1 y *Accuracy*. Plantean una posible línea de trabajo futura que combine el uso de esta arquitectura con el algoritmo de representación de texto que presentaron en la anterior edición del TASS. La implementación del sistema la realizaron en Python, haciendo uso de la librería Tensorflow con soporte para GPU.

En Cárdenas et al. [7] proponen un algoritmo para la clasificación automática de textos, como una alternativa a los tradicionalmente utilizados en esta tarea. El clasificador propuesto considera la dependencia entre las variables predictoras (palabras o términos), algo que los clasificadores de texto comúnmente utilizados no hacen. La dependencia entre estas variables queda plasmada en forma de enlaces en grafos de palabras co-ocurrentes, objetos utilizados para entrenar el clasificador y además estimar la categoría de un texto desconocido. Los resultados obtenidos al clasificar automáticamente el sentido positivo, negativo o neutral de más de 1,000 mensajes de Twitter escritos en español, en distintos contextos (temas), muestran que el algoritmo, además de ser una propuesta novedosa para la clasificación automática de textos, tiene un desempeño, al menos, similar al de otros tradicionalmente utilizados en este tipo de problemas, como las Máquinas de Soporte Vectorial o algoritmos de estadística Bayesiana. Para evaluar el desempeño del clasificador se utilizaron las tradicionales medidas de precisión (exactitud) y cobertura respecto a la clasificación realizada por un humano.

En Escolano y Costa-jussà [9] plantean dividir la tarea de traducción en dos partes: primero, simplificaron el lenguaje destino en términos morfológicos y cons-

truyeron el sistema de traducción con esta modificación; y después utilizaron un algoritmo de clasificación (bayesiano ingenuo, SVMs, Random Forest, etc.) para generar la morfología final. El corpus utilizado consiste en fragmentos extraídos de discursos de la ONU. Para cada uno de los fragmentos dispusieron de sus correspondientes traducciones en chino y castellano. El trabajo presenta una arquitectura de aprendizaje profundo que permite añadir de manera efectiva la información morfológica a la traducción simplificada generada por un traductor estadístico basado en segmentos. Donde demostraron que la arquitectura diseñada presenta resultados superiores que los presentados en el estado del arte en términos de precisión y calidad de la traducción mejora en términos de METEOR.

En la tesis presentada por Ángel Javier Alonso Hernández [19] se estudia la aplicación del *deep learning* a la tarea del resumen automático y abstractivo de textos. El conjunto de datos con el que se trabajó se llama CNN-DailyMail, que consiste en aproximadamente 300 000 pares de artículos periodísticos y su resumen se analizaron las técnicas *deep learning* enfocadas al procesamiento de lenguaje natural: las redes neuronales recurrentes y los modelos *encoder-decoder*, entre otras. Para ello, se realizó un estudio de la bibliografía y, siguiendo las últimas líneas de investigación, donde se diseñó un modelo propio, que cuenta con una arquitectura *encoder-decoder* y un mecanismo de atención. Este modelo se entrenó sobre una tarjeta gráfica donada por NVIDIA. Finalmente, evaluaron los resultados con las métricas ROUGE-1, ROUGE-2 y ROUGE-L. Aunque en los resultados que obtuvieron, muchos de los resúmenes son buenos, se identificaron algunos problemas como la repetición de frases y la falta de vocabulario.

En Carrera [6] presenta una investigación la cual tiene como objetivo clasificar los sílabos mediante la técnica de redes neuronales conectadas de aprendizaje profundo, a través de una combinación de numero de capas, funciones de activación, tamaños y épocas de entrenamiento. El conjunto de datos con el que se realizó la experimentación consta de 2,316 sílabos provenientes de la universidad del Azuay y la Universidad de Cuenca divididos en nueve clases y 21 subclases. El lenguaje de programación utilizado fue Python. El modelo fue comparado con respecto a algoritmos basados máquinas de soporte vectorial (SVM), Naïve Bayes y árboles de decisión, realizando la evaluación con la medida de exactitud. Los resultados demostraron que

el modelo de aprendizaje profundo propuesto fue superior en 1.4 % con respecto a Naïve Bayes, 6.2 % con respecto a SVM y 7.2 % con respecto a árboles de decisión.

En la tesis presentada por Armijos [4] propone consolidar y preparar un cuerpo con expresiones de texto extraídas de Twitter el cual permite analizar la información a través de la ciencia de datos, con el fin de utilizarlo como insumo esencial para entrenar una red neuronal convolucional (CNN), mediante técnicas de aprendizaje profundo. Como resultado de este entrenamiento, se genera un modelo de predicción de textos que puedan o no presentar signos de cyber acoso (cyber bullying), en los cuales se pueda manifestar agresión verbal grave, como insultos, ataques racistas, ataques homofóbicos, etc. El modelo fue validado mediante técnicas de validación cruzada, obteniendo resultados satisfactorios que permiten concluir que el modelo generado es óptimo para realizar predicciones de textos para la identificación del ciberacoso.

En el libro escrito por Varga [34], se menciona lo que es la ciencia de datos y las etapas del ciclo de vida, las cuales son: iniciación del proyecto, adquisición de datos, preparación de datos, análisis de datos, informes y ejecución de acciones. Se mencionan las etapas a detalle. También hace mención de lo que es *deep learning*, *machine learning*, redes neuronales, visualización de datos, seguridad de datos y como hacer la documentación de un proyecto. También se presentan los datos más útiles de Python 3 como marcos y herramientas científicas: Numpy, Pandas, scikit-learn, matplotlib, Seaborn, Dask, Apache Spark, PyTorch y otros marcos auxiliares. Al final de cada capítulo se presenta un ejercicio para dar un mejor aprendizaje. El libro es muy útil puesto que presenta las fases de ciencia de datos, así como varios ejemplos donde se utiliza el lenguaje de programación Python.

Capítulo 3

Marco teórico

En este capítulo, se presentan y explican los conceptos teóricos usados en el presente trabajo de tesis.

3.1. Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN) según Vásquez et al. [35] es una rama de la inteligencia artificial la cual se centra en la utilización del Lenguaje Natural (LN), para mantener una comunicación entre las personas y las computadoras. Con el lenguaje natural se facilita la creación de programas o modelos que nos ayudan a entender el manejo del lenguaje natural.

3.1.1. Arquitectura de un sistema de PLN

La arquitectura de un sistema de PLN se sustenta en una definición del LN por niveles, cada nivel no es independiente del anterior pues existe una relación entre cada uno de ellos, por lo cual si se requiere analizar un nivel se necesita conocer de forma básica los niveles anteriores a este(ver Figura 3.1), los niveles son los siguientes:

- Nivel Fonológico: trata de cómo las palabras se relacionan con los sonidos que representan.
- Nivel Morfológico: trata de cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.

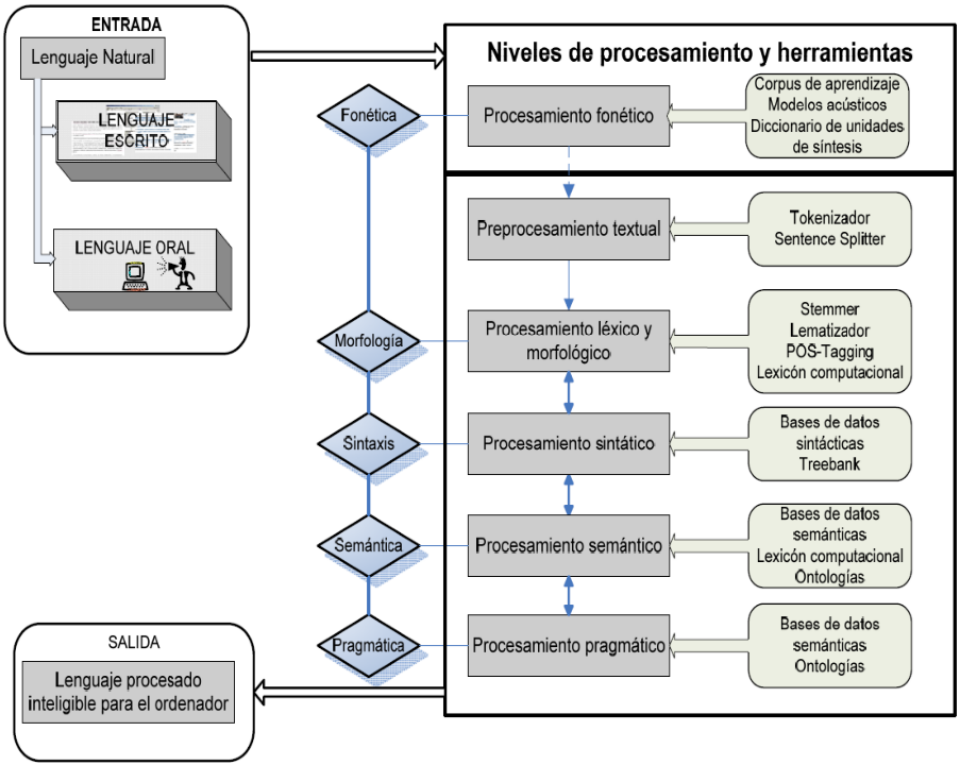

Figura 3.1: Niveles de Procesamiento de Lenguaje Natural [10].

- Nivel Sintáctico: trata de cómo las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.

- Nivel Semántico: trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.

- Nivel Pragmático: trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

3.1.2. Aplicaciones del PLN

Las aplicaciones del PLN son muy variadas, ya que su alcance es muy grande y ha dado lugar a múltiples líneas de investigación, algunas de las aplicaciones son:

- Traducción automática
- Recuperación de la información
- Extracción de Información y Resúmenes
- Resolución cooperativa de problemas
- Tutores inteligentes
- Reconocimiento de Voz

3.2. Recuperación de la información

La complejidad asociada al procesamiento de lenguaje natural [33] cobra especial relevancia cuando necesitamos recuperar información textual que satisfaga la necesidad de información de un usuario. Es por ello, que en el área de Recuperación de Información Textual las técnicas de PLN son muy utilizadas, tanto para facilitar la descripción del contenido de los documentos, como para representar la consulta formulada por el usuario, y ello, con el objetivo de comparar ambas descripciones y presentar al usuario aquellos documentos que satisfagan en mayor grado su necesidad de información.

Dicho de otro modo, un sistema de recuperación de información textual lleva a cabo las siguientes tareas para responder a las consultas de un usuario (ver Figura 3.2):

1. Indexación de la colección de documentos: en esta fase, mediante la aplicación de técnicas de PLN, se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento es descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.

2. Cuando un usuario formula una consulta el sistema la analiza, y si es necesario la transforma, con el fin de representar la necesidad de información del usuario del mismo modo que el contenido de los documentos.

3. El sistema compara la descripción de cada documento con la descripción de la consulta, y presenta al usuario aquellos documentos cuyas descripciones más se asemejan a la descripción de su consulta.

4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.

3.2.1. Campos de investigación relacionados

Existen diferentes campos de investigación relacionados con la recuperación de información y el procesamiento del lenguaje natural que enfocan el problema desde otra perspectiva, pero cuyo objetivo final es facilitar el acceso a la información.

- La extracción de información consiste en extraer las entidades, los eventos y relaciones existentes entre los elementos de un texto o de un conjunto de textos. Es una forma de acceder eficientemente a documentos grandes, pues extrae partes del documento que muestran el contenido de éste. La información generada puede utilizarse para bases de conocimiento u ontologías.

- La generación de resúmenes se basa en condensar la información más relevante de un texto. Las técnicas utilizadas varían según la tasa de compresión, la finalidad del resumen, el género del texto, el idioma (o idiomas) de los textos de partida, entre otros factores.

- La búsqueda de respuesta tiene como objetivo dar una respuesta concreta a la pregunta formulada por el usuario. Las necesidades de información han de estar muy definidas: fechas, lugares, etc. En este caso el procesamiento del lenguaje natural trata de identificar el tipo de respuesta a facilitar (mediante la desambiguación de la pregunta, el análisis de las restricciones fijadas, y el

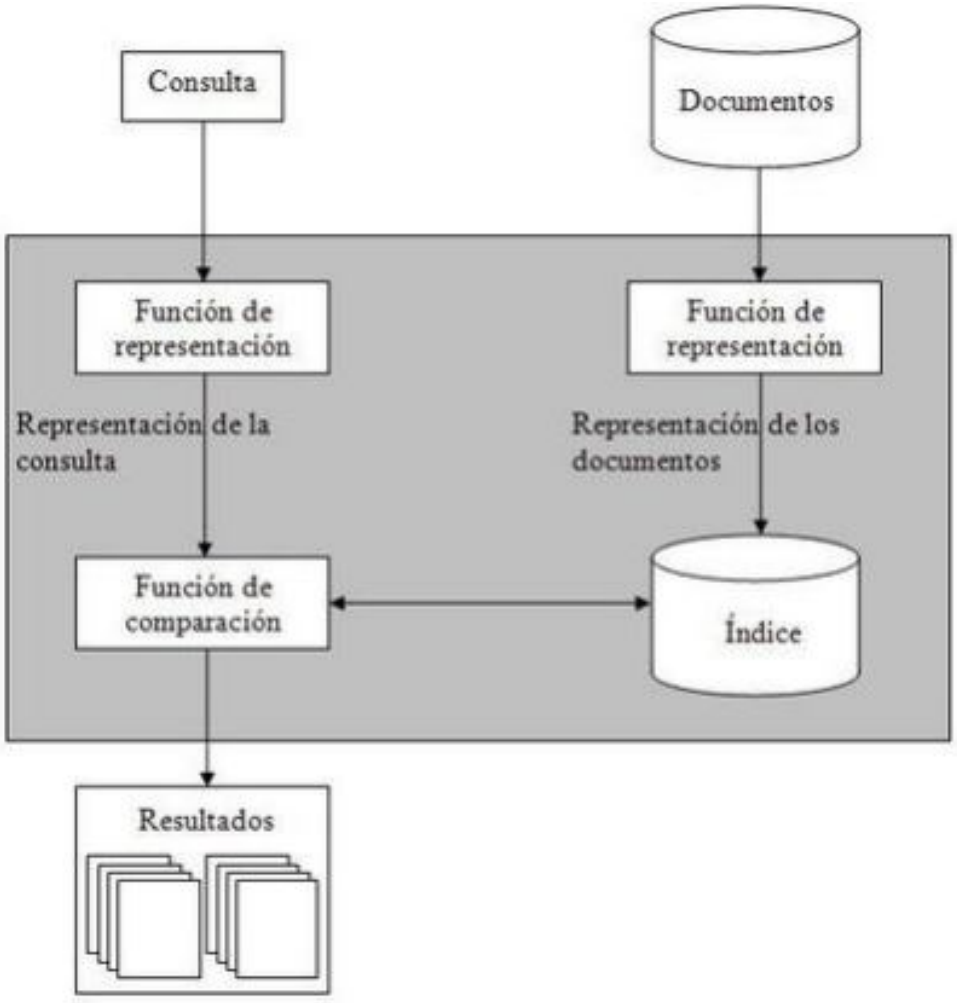

Figura 3.2: Arquitectura de un sistema de recuperación de información [33].

uso de técnicas para la extracción de información). Estos sistemas son considerados como uno de los potenciales sucesores de los actuales sistemas de recuperación de información. START natural language system es un ejemplo de estos sistemas.

- La recuperación de información multilingüe consiste en la posibilidad de recuperar información aunque la pregunta y/o los documentos estén en diferentes idiomas. Son utilizados traductores automáticos de los documentos y/o de las preguntas, o mecanismos interlingua para crear interpretaciones de los docu-

mentos. Estos sistemas suponen un gran reto, pues combinan dos aspectos claves en el actual contexto de la web, la recuperación de información y el tratamiento de información multilingüe.

- Para acabar, hay que citar las técnicas automáticas de clasificación de textos, consistentes en la asignación automática de un conjunto de documentos a categorías de clasificación predefinidas. La correcta descripción de las características de los documentos (normalmente mediante el uso de técnicas estadísticas - preprocesado y parametrización) determinará en gran medida la calidad de los agrupamientos/categorizaciones propuestos por estas técnicas.

3.2.2. Reconocimiento de entidades con nombre

El reconocimiento de entidades con nombre según Nadeau y Sekine [18] es una tarea importante de la recuperación de información y fue llamado "Reconocimiento y clasificación de entidades con nombre (*Named Entity Recognition and Classification*, NERC)". La cual consiste en reconocer las unidades de información como nombres, incluyendo a la persona, organización y nombres de lugares y expresiones numéricas incluyendo hora, fecha, el dinero y el porcentaje de expresiones. Las investigaciones mayormente se concentran en el idioma inglés, pero a lo largo del tiempo han surgido más idiomas tales como: Alemán, Español, Chino, etc.

Algunos tipos de entidad son: Enamex: nombres propios: los nombres de personas, lugares y organizaciones GPE: lugar que tiene un gobierno, tal como una ciudad o de un país. Miscellaneous: nombres propios que quedan fuera del enamex clásico Timex: tipos fecha y tiempo y el dinero Numex: tipo moneda y porcentaje.

Estos son algunos de los tipos de entidades con los que se comenzó, hoy en día se han surgido muchas variantes dependiendo a los diferentes intereses. El número de categorías consta de unas 200, las cuales son las entidades más utilizadas. A continuación se muestra un ejemplo en la Figura 3.3, en donde se puede observar los siguientes valores:

- LOC: representado de color amarrillo, se refiere a lugares, cadenas montañosas, cuerpos de agua; en este caso se mencionan: Puerto Escondido, Oaxaca, Punta Colorada, La Fiscalía, Estado de Oaxaca y FISCALIA_GobOAX.

- PER: representado de color gris, se refiere a las personas, en este caso se mencionan: María Eugenia.
- ORG: representado de color verde, se refiere a las empresas, agencias, instituciones, etc; en este caso se mencionan: Fiscalía de Oaxaca.
- MISC: representado de color azul, se refiere a entidades diversas, por ejemplo, eventos, nacionalidades, productos u obras de arte; en este caso se mencionan: Comandancia Regional de la Policía Estatal, el cuerpo y “.

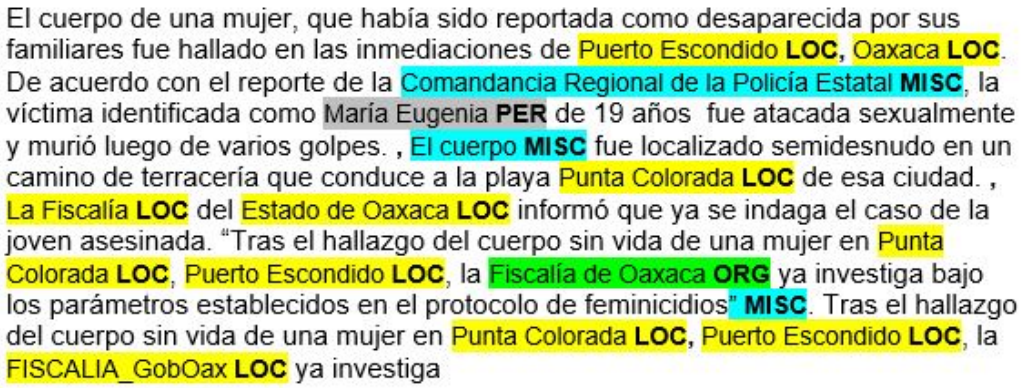
El cuerpo de una mujer, que había sido reportada como desaparecida por sus familiares fue hallado en las inmediaciones de Puerto Escondido **LOC**, Oaxaca **LOC**. De acuerdo con el reporte de la Comandancia Regional de la Policía Estatal **MISC**, la víctima identificada como María Eugenia **PER** de 19 años fue atacada sexualmente y murió luego de varios golpes. , El cuerpo **MISC** fue localizado semidesnudo en un camino de terracería que conduce a la playa Punta Colorada **LOC** de esa ciudad. , La Fiscalía **LOC** del Estado de Oaxaca **LOC** informó que ya se indaga el caso de la joven asesinada. “Tras el hallazgo del cuerpo sin vida de una mujer en Punta Colorada **LOC**, Puerto Escondido **LOC**, la Fiscalía de Oaxaca **ORG** ya investiga bajo los parámetros establecidos en el protocolo de feminicidios” **MISC**. Tras el hallazgo del cuerpo sin vida de una mujer en Punta Colorada **LOC**, Puerto Escondido **LOC**, la FISCALIA_GobOax **LOC** ya investiga

Figura 3.3: Ejemplo de reconocimiento de entidades con nombre.

3.3. Ciencia de datos

La ciencia de datos segun Varga [34] puede describirse como la aplicación de principios y métodos científicos a la recopilación, análisis e información de datos. El objetivo es sintetizar información confiable y procesable a partir de datos. El proceso es inherentemente iterativo e incremental. El proceso de ciencia de datos, incluye seis fases o actividades (ver Figura 3.4) distintas que dependen unas de otras, el proceso es iterativo y los resultados de un paso pueden requerir que el paso anterior se repita con nueva información.

A continuación se presenta un resumen de cada fase:

1. Iniciación del proyecto: indicar el problema que está tratando de resolver o la pregunta que está tratando de responder. Se debe comenzar por comprender

los objetivos, antes de comenzar a pensar en soluciones, se debe trabajar con ellos para definir claramente el problema.

2. Adquisición de los datos: adquirir incluye todo lo que nos hace recuperar datos, incluidos; encontrar, acceder, adquirir y mover datos. Incluye identificación y acceso autenticado a todos los datos relacionados.

3. Preparación de los datos: se divide la fase previa en: explorar datos y preprocesar datos. El primer paso en la preparación de datos consiste literalmente en mirar los datos para comprender su naturaleza, lo que significa, su calidad y formato. A menudo se necesita un análisis preliminar de datos, o muestras de datos, para comprenderlo. Es por eso que este paso se llama explorar. Una vez que sepamos más sobre los datos a través del análisis exploratorio, el siguiente paso es el preprocesamiento de los datos para el análisis. El preprocesamiento incluye datos de limpieza, subconfiguración o filtrado de datos, creación de datos que los programas pueden leer y comprender, como modelar datos en bruto en un modelo de datos más definido o empaquetarlos con un formato de datos específico. Si hay varios conjuntos de datos involucrados, este paso también incluye la integración de múltiples fuentes de datos o secuencias.

4. Análisis de datos: Los datos preparados se pasarían al paso de análisis, que implica la selección de técnicas analíticas para usar, construir un modelo de los datos y analizar los resultados. Esta fase puede tomar un par de iteraciones por sí sola o puede requerir que los científicos de datos vuelvan a los pasos uno y dos para obtener más datos o datos de paquetes de una manera diferente.

5. Informes: para comunicar los resultados se incluye la evaluación de los resultados analíticos. Presentarlos de manera visual, crear informes que incluyen una evaluación de resultados con respecto a los criterios de éxito. Las actividades en esta fase a menudo se pueden referir con términos como interpretar, resumir, visualizar o publicar el proceso.

6. Ejecución de acciones: esta última fase nos lleva de vuelta a la primera razón por la que se hace ciencia de datos, el propósito. Informar los conocimientos

del análisis y determinar las acciones a partir de los conocimientos basados en el propósito que se definió inicialmente es lo que se llama la fase del acto.

Hay tres áreas de competencias centrales interconectadas en ciencia de datos: conocimiento de dominio, matemáticas (incluyendo teoría de probabilidad y estadística) e ingeniería de software (una sola persona puede asumir diferentes roles en distintos momentos). Otro denominador es la automatización (a través de la capacidad de programación de la mayoría de las actividades) para aumentar la productividad, la reproducibilidad y la calidad. El objetivo de aprender ciencia de datos significa obtener el conocimiento y la comprensión de lo que conlleva la aplicación de ciencia de datos a través de lo estudiado y la experiencia obtenida.

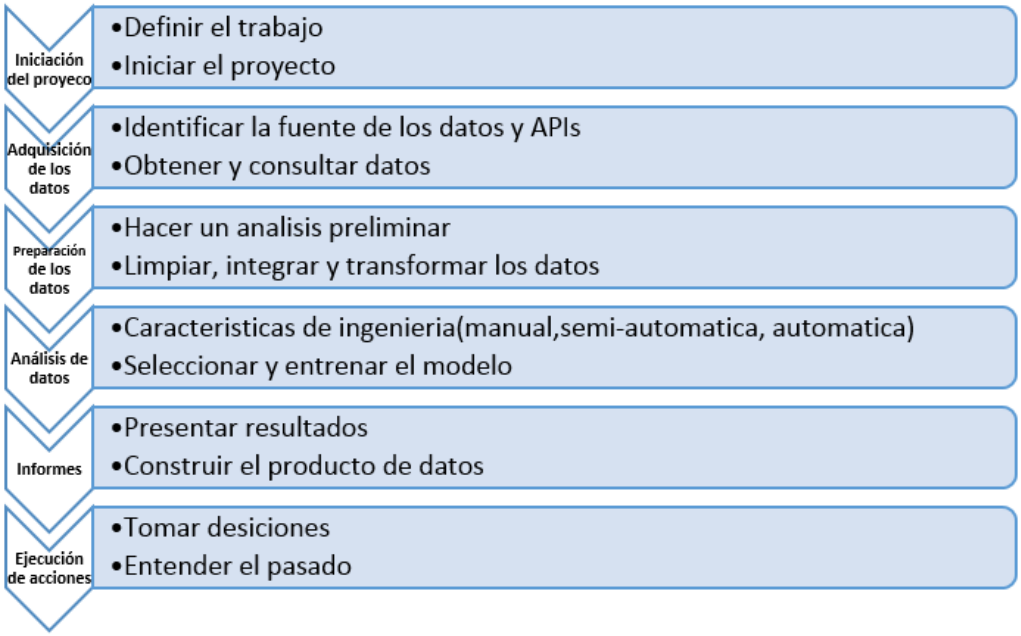

Figura 3.4: Las principales fases de un proyecto de ciencia de datos [34].

3.4. Aprendizaje automático

El aprendizaje automático o *machine learning* es un subcampo de las ciencias computacionales y una rama de la inteligencia artificial, cuyo objetivo se encarga de desarrollar algoritmos o técnicas que tiene la capacidad de aprender, para resolver

problemas de diversos campos. Para que el algoritmo tenga la capacidad de aprender solo basta con alimentarlo con una gran cantidad de datos y este sea capaz de saber qué hacer en cada uno de los casos posibles (Sandoval [26]).

En este contexto, se dice que un sistema que aprende de forma automatizada (o aprendiz) es un artefacto (o un conjunto de algoritmos) que, para resolver problemas, toma decisiones basadas en la experiencia acumulada (en los casos resueltos anteriormente) para mejorar su actuación. Estos sistemas deben ser capaces de trabajar con un rango muy amplio de tipos de datos de entrada, que pueden incluir datos incompletos, inciertos, ruido, inconsistencias, etc. Nuestra primera caracterización del proceso de aprendizaje automático es: *Aprendizaje = Selección + Adaptación* visto así, el aprendizaje automático es un proceso que tiene lugar en dos fases. Una en la que el sistema elige (selecciona) las características más relevantes de un objeto (o un evento), las compara con otras conocidas - si existen- a través de algún proceso de cotejamiento y, cuando las diferencias son significativas, adapta su modelo de aquel objeto (o evento) según el resultado del cotejamiento, Moreno et al. [17].

3.4.1. Tipos de aprendizajes

Hay tres tipos de aprendizajes: el supervisado, el no supervisado y mediante refuerzos.

1. Aprendizaje supervisado: Es cuando entrenamos un algoritmo de aprendizaje automático dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro el algoritmo pueda hacer una predicción conociendo las características. En este tipo de aprendizaje hay dos algoritmos (entrenamientos): el de clasificación y el de regresión.

    - Algoritmo de clasificación: esperamos que el algoritmo nos diga a qué grupo pertenece el elemento en estudio. El algoritmo encuentra patrones en los datos que le damos y los clasifica en grupos. Luego compara los nuevos datos y los ubica en uno de los grupos y es así como puede predecir de que se trata.

      La variable por predecir es un conjunto de estados discretos o categóricos. Pueden ser:

Binaria: Sí, No, Azul, Rojo, Fuga, No Fuga, etc.

Múltiple: Comprará Producto1, Producto 2..., etc.

Ordenada: Riesgo Bajo, Medio, Alto, etc

- Algoritmo de regresión: en este método lo que se espera es un número. No lo ubica en un grupo, sino que devuelve un valor específico.

2. Aprendizaje no supervisado: Aquí solo le damos las características al algoritmo, nunca las categorias. Queremos que nos agrupe los datos que le dimos según sus características. El algoritmo solo sabe que los datos comparten ciertas características, de esa forma asume que pueda que pertenezcan al mismo grupo.

3. Mediante refuerzos: Este método de aprendizaje está a medio camino entre los dos anteriores. Al sistema se le proponen problemas que debe solucionar. El aprendizaje se realiza únicamente con una señal de refuerzo proporcionada por un profesor o por el entorno como indicador de si se ha resuelto correctamente el problema.

3.4.2. Modelos de aprendizaje automático

Los algoritmos de aprendizaje automático, se pueden agrupar en tres modelos:

- Modelos lineales: Estos tratan de encontrar una línea que se "ajuste" bien a la nube de puntos que se disponen. Aquí destacan desde modelos muy conocidos y usados como la regresión lineal (también conocida como la regresión de mínimos cuadrados), la logística (adaptación de la lineal a problemas de clasificación - cuando son variables discretas o categóricas-). Estos dos modelos tienen el problema del *"overfit"*, esto significa que se ajustan "demasiado" a los datos disponibles, con el riesgo que esto tiene para nuevos datos que pudieran llegar. Al ser modelos relativamente simples, no ofrecen resultados muy buenos para comportamientos más complicados.

- Modelos de árbol: Son modelos precisos, estables y más sencillos de interpretar básicamente porque construyen unas reglas de decisión que se pueden representar como un árbol. A diferencia de los modelos lineales, pueden representar

relaciones no lineales para resolver problemas. En estos modelos, destacan los árboles de decisión y los *random forest* (una media de árboles de decisión). Al ser más precisos y elaborados, obviamente ganamos en capacidad predictiva, pero perdemos en rendimiento.

- Redes neuronales: Las redes artificiales de neuronas tratan, en cierto modo, de replicar el comportamiento del cerebro, donde tenemos millones de neuronas que se interconectan en red para enviarse mensajes unas a otras. Esta réplica del funcionamiento del cerebro humano es uno de los "modelos de moda" por las habilidades cognitivas de razonamiento que adquieren. El reconocimiento de imágenes o vídeos, por ejemplo, es un mecanismo complejo y una red neuronal es lo mejor para realizarlo. El problema, como ocurre con el cerebro humano, es que son lentas de entrenar y necesitan mucha capacidad de cómputo. Quizás sea uno de los modelos que más ha ganado con la "revolución de los datos".

3.4.3. Modelos de aprendizaje automático más utilizados

Algunos de los tantos modelos con los que cuenta el aprendizaje automático, son los siguiente cuatro que se presentan a continuación Sampedro [25]:

- Naïve Bayes

  Naïve Bayes se define como un modelo probabilístico que debe su nombre a dos circunstancias: aplicación del Teorema de Bayes para extraer la regla de clasificación y la simplicidad de suponer que todos los atributos son condicionalmente independientes dada la clase (naïve = ingenuo, simple). Dicha suposición está bastante alejada de la realidad en la mayoría de problemas. No obstante, este modelo ofrece un rendimiento sorprendentemente bueno debido a que no hay que estimar distribuciones de probabilidad conjuntas, evitando así sufrir la maldición de la dimensionalidad.

  La función de clasificación del método Naïve Bayes se define basándose en la regla del Máximo A Posteriori (MAP) de las probabilidades de cada clase:

$$f(I) = argmax_k \mathbb{P}(\omega(I) = C_k|I) = argmax_k \mathbb{P}(\omega(I) = C_k|A_1, ..., A_a). \quad (3.1)$$

Desarrollando la expresión anterior mediante el Teorema de Bayes y aplicando la suposición de independencia se obtiene:

$$f(I) = argmax_k \mathbb{P}(C_k) \prod_{j=1}^{a} \mathbb{P}(\Psi(A_j, I)|C_k). \tag{3.2}$$

Los términos $\mathbb{P}(C_k)$ y $\mathbb{P}(\Psi(A_j, I)|C_k)$ representan las probabilidades a priori de que una instancia I sea de la clase C_k y de que el atributo A_j tome cierto valor para las instancias de la clase C_k, respectivamente. Se pueden calcular con facilidad a partir de los datos de entrenamiento, haciendo un recuento del número de instancias de cada clase y, entre ellas, el número de veces que aparece cada valor del atributo. Finalmente, se divide cada cantidad entre el total de casos posibles (total de instancias para $\mathbb{P}(C_k)$ y total de instancias de la clase C_k para $\mathbb{P}(\Psi(A_j, I)|C_k)$. Si los atributos son continuos se pueden discretizar y tratarlos como tal, o bien ajustar los parámetros de una distribución conocida, típicamente una Gaussiana, para estimar las probabilidades buscadas. La primera opción es más recomendable si se dispone de gran cantidad de datos.

- Vecinos más cercanos

  Los vecinos más cercanos se definen como un modelo perteneciente al paradigma de aprendizaje perezoso, que se caracteriza por la ausencia de una fase de extracción de conocimiento de los datos. El funcionamiento está dirigido por la demanda de clasificación de una nueva instancia. En ese momento se analiza el entorno de dicha instancia y se emite una predicción basada, como el propio nombre del algoritmo indica, en las instancias vecinas más cercanas a la demandada. El modelo se puede aplicar tanto a problemas de clasificación como de regresión. El algoritmo de predicción es muy sencillo. En primer lugar se fija un natural que determina el número de vecinos que se tendrán en cuenta y se hallan la b instancias más cercanas a una dada, empleando alguna distancia (euclídea, Mahalanobis, etc.). A continuación, si se trata de un problema de clasificación, se hace un recuento de cuántos de los $b - vecinos$ pertenecen a cada clase y se asigna a la nueva instancia la clase mayoritaria; si se trata de un problema de regresión, se calcula el promedio de los valores de las etiquetas de los b-vecinos y se asigna el resultado a la etiqueta de la nueva instancia. En

ambos casos es posible ponderar la influencia de cada vecino con una función dependiente de la distancia, de m.

- Redes neuronales

  Inicialmente, las redes neuronales surgieron por la motivación de tratar de reproducir el funcionamiento del cerebro. En 1943, McCulloch y Pitts elaboraron un modelo artificial de una neurona, que simulaba el procesamiento de información que tenía lugar en una neurona real. A partir de entonces, se ha tratado de construir redes cada vez más complejas pero, de momento, el objetivo de emular el comportamiento del cerebro está lejos de alcanzarse. Uno de los primeros modelos fue el Perceptrón de Rosenblatt, que permite establecer un hiperplano en un espacio a-dimensional mediante el ajuste de los parámetros (= pesos sinápticos) de las neuronas artificiales. Con este modelo se pueden resolver problemas lineales. Sin embargo, como la mayoría de problemas reales no son lineales, se desarrolló más este enfoque incluyendo varias capas de neuronas artificiales y generando un nuevo algoritmo de aprendizaje (*Backpropagation*) para ajustar los pesos de la red. El aprendizaje se basa en descenso por gradiente del error de predicción. Este nuevo modelo se denomina Perceptrón Multicapa y se demostró que es una familia de aproximación universal, lo que significa que puede aproximar cualquier función continua con tanta precisión como se desee (siempre que se disponga de un número suficiente de instancias).

- Máquinas de vectores de soporte

  Este modelo es parecido a las redes neuronales en cuanto a su objetivo de ajustar un conjunto de parámetros, que permiten establecer fronteras en el espacio a-dimensional y aproximar funciones o separar patrones en diferentes regiones del espacio de atributos. La diferencia radica en el método de entrenamiento para ajustar los parámetros. Las máquinas de vectores de soporte basan su entrenamiento en la maximización del margen existente entre el hiperplano separador y las instancias de las dos clases (inicialmente, este modelo se diseñó para resolver problemas de clasificación de dos clases pero hay extensiones para problemas multi-clase y para problemas de regresión). Las máquinas de vec-

tores de soporte se comportan de manera muy similar a las redes neuronales. Presentan un rendimiento muy bueno pero el hecho de que el entrenamiento consista en el ajuste de unos parámetros ocultos y la predicción consista únicamente en un resultado numérico, hace que tanto la transparencia como la capacidad de explicación sean muy pobres.

3.5. Aprendizaje profundo

Algunas de las definiciones que tiene el aprendizaje profundo o *deep learning* (DL), son las siguientes Deng y Yu [8]:

1. Una clase de técnicas de aprendizaje automático que explotan muchas capas de procesamiento de información no lineal para la extracción y transformación de características supervisadas o no supervisadas, y para el análisis y clasificación de patrones.

2. "Un subcampo dentro del aprendizaje automático que se basa en algoritmos para aprender múltiples niveles de representación para modelar relaciones complejas entre datos. Las características y conceptos de nivel superior se definen así en términos de los de nivel inferior, y dicha jerarquía de características se denomina arquitectura profunda. La mayoría de estos modelos se basan en el aprendizaje no supervisado de representaciones ".

3. "Un subcampo de aprendizaje automático que se basa en el aprendizaje de varios niveles de representaciones, correspondientes a una jerarquía de características o factores o conceptos, donde los conceptos de nivel superior se definen a partir de los de nivel inferior, y el mismo nivel inferior Los conceptos pueden ayudar a definir muchos conceptos de nivel superior. El aprendizaje profundo es parte de una familia más amplia de métodos de aprendizaje automático basados en representaciones de aprendizaje. Una observación (Por ejemplo: Una imagen) se puede representar de muchas maneras (Por ejemplo: Un vector de píxeles), pero algunas representaciones facilitan el aprendizaje de tareas de interés (Por ejemplo: ¿Es está la imagen de un rostro humano?)

A partir de ejemplos, y la investigación en esta área intenta definir qué hace mejores representaciones y cómo aprenderlas".

4. "El aprendizaje profundo es un conjunto de algoritmos en el aprendizaje automático que intentan aprender en múltiples niveles, correspondientes a diferentes niveles de abstracción. Por lo general, utiliza redes neuronales artificiales. Los niveles en estos modelos estadísticos aprendidos corresponden a distintos niveles de conceptos, donde los conceptos de nivel superior se definen a partir de los de nivel inferior, y los mismos conceptos de nivel inferior pueden ayudar a definir muchos conceptos de nivel superior ".

La base del aprendizaje profundo son las redes neuronales descritas por [36] [37], que se combinan formando las redes neuronales profundas. Estas técnicas han demostrado éxitos empíricos del aprendizaje profundo en diversas aplicaciones de visión por computadora, en los campos del procesamiento de sonido e imagen, incluido el reconocimiento facial, el reconocimiento de voz, el procesamiento automatizado del lenguaje, la clasificación de texto (por ejemplo, el reconocimiento de spam), recuperación de información, robótica e incluso en el análisis de moléculas que pueden conducir al descubrimiento de nuevos fármacos. Las posibles aplicaciones son muy numerosas. Un ejemplo espectacular es el programa AlphaGo, que aprendió a jugar con el método de aprendizaje profundo y derrotó al campeón mundial en 2016.

Existen varios tipos de arquitecturas para redes neuronales (ver Figura 3.5):

- Los perceptrones multicapa, que son los más antiguos y simples.
- Las redes neuronales convolucionales ("*Convolutional Neural Networks*", CNN), especialmente adaptadas para el procesamiento de imágenes.
- Las redes neuronales recurrentes, utilizadas para datos secuenciales como texto o series de tiempo.

Las redes neuronales se basan en una profunda cascada de capas. Necesitan algoritmos inteligentes de optimización estocástica e inicialización, y también una elección inteligente de la estructura.

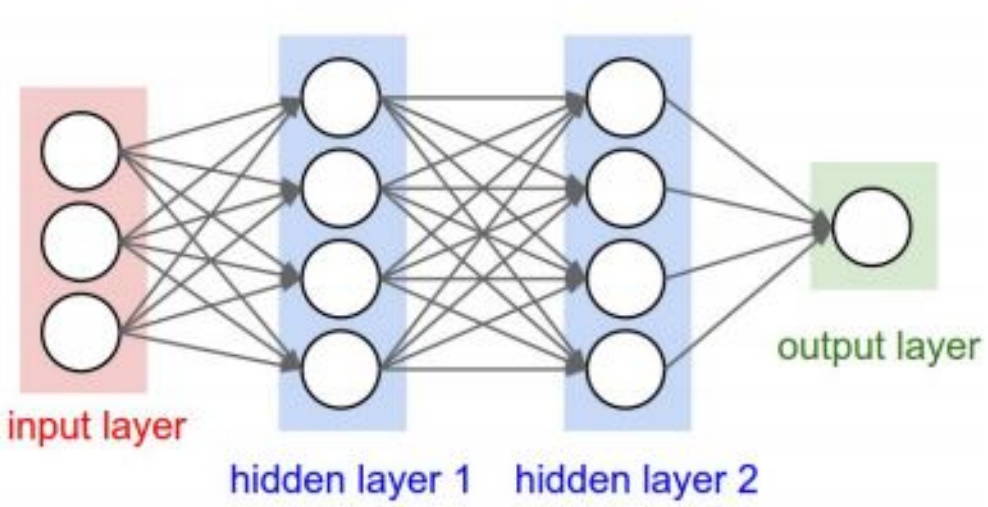

Figura 3.5: Una red neuronal básica [36].

3.5.1. Redes neuronales convolucionales

El nombre "red neuronal convolucional (CNN)" indica que la red emplea una operación matemática llamada convolución. La convolución es un tipo especializado de operación lineal. Las redes convolucionales son simplemente redes neuronales que utilizan la convolución en lugar de la multiplicación matricial general en al menos una de sus capas. CNN es un tipo de modelo de aprendizaje profundo para procesar datos que tiene un patrón de cuadrícula, como imágenes, que está inspirado en la organización de la corteza visual animal y diseñado para aprender de forma automática y adaptativa las jerarquías espaciales de las características, desde patrones de bajo a alto nivel.

CNN es una construcción matemática que generalmente se compone de tres tipos de capas (o bloques de construcción): convolución, agrupación y capas completamente conectadas, las cuales se describen a continuación.

- Capa de convolución

  La convolución discreta entre dos funciones f y g se define como

$$(f \cdot g)(x) = \sum_t f(t)g(x+t). \tag{3.3}$$

Para señales bidimensionales como imágenes, consideramos las convoluciones 2D

$$(K \cdot I)(i,j) = \sum_{m,n} K(m,n) I(i+n, j+m). \tag{3.4}$$

K es un núcleo de convolución aplicado a una señal 2D (o imagen) I.

La convolución es un tipo especializado de operación lineal utilizada para la extracción de características, donde se aplica una pequeña matriz de números, llamada núcleo, a través de la entrada, que es una matriz de números, llamada tensor. Se calcula un producto basado en elementos entre cada elemento del núcleo y el tensor de entrada en cada ubicación del tensor y se suma para obtener el valor de salida en la posición correspondiente del tensor de salida, denominado mapa de características (ver Figura 3.6). La operación de convo-

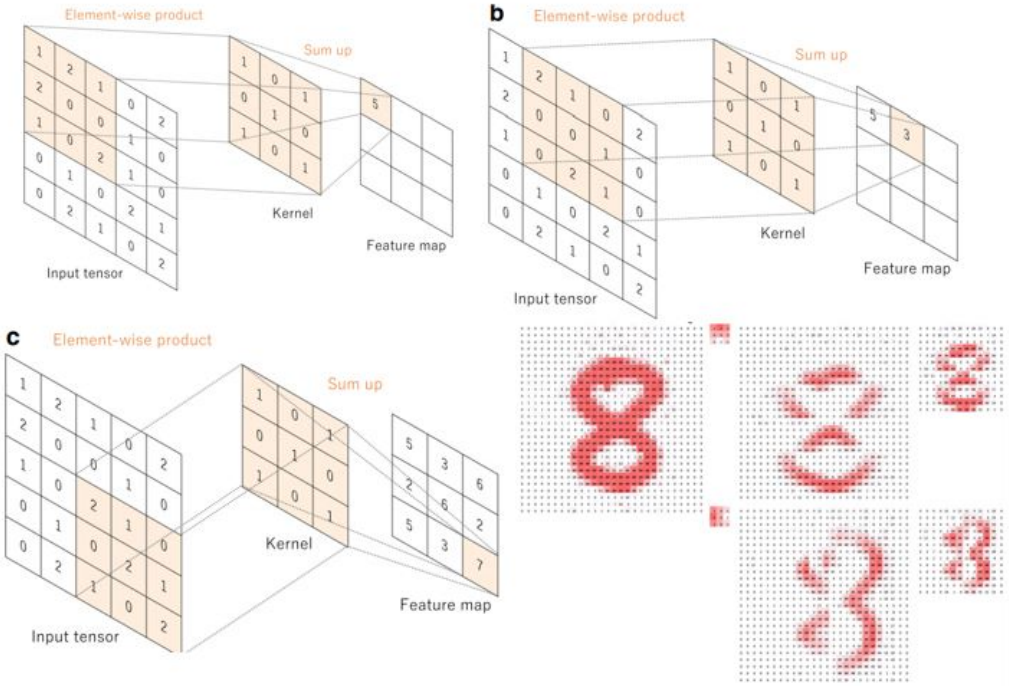

Figura 3.6: Un ejemplo de operación de convolución con un tamaño de núcleo de 3×3 [37].

lución no permite que el centro de cada núcleo se superponga con el elemento más externo del tensor de entrada, y reduce la altura y el ancho del mapa de

Tabla 3.1: Una lista de las funciones de activación de la última capa comúnmente aplicadas para diversas tareas [37].

	Parámetros	Hiperparámetros
Capa de convolución	Granos	Tamaño del grano, número de granos, zancada, relleno, función de activación
Capa de agrupación	Ninguna	Método de agrupación, tamaño de filtro, zancada, relleno
Capa completamente conectada	Pesas	Número de pesas, función de activación
Otros		Arquitectura del modelo, optimizador, tasa de aprendizaje, función de pérdida, tamaño de mini lote, épocas, regularización, inicialización de peso, división de conjuntos de datos

características de salida en comparación con el tensor de entrada. El relleno, generalmente llamado relleno cero, es una técnica para abordar este problema, donde se agregan filas y columnas de ceros a cada lado del tensor de entrada, para ajustar el centro de un núcleo en el elemento más externo y mantener el mismo plano dimensión a través de la operación de convolución (ver Figura 3.7). La distancia entre dos posiciones sucesivas del núcleo se llama zancada, que también define la operación de convolución. La elección común de un paso es 1; sin embargo, a veces se usa una zancada mayor que 1 para lograr la disminución de la resolución de los mapas de características. Una técnica alternativa para realizar el muestreo descendente es una operación de agrupación. El proceso de capacitación de un modelo CNN con respecto a la capa de convolución es identificar los núcleos que funcionan mejor para una tarea determinada en función de un conjunto de datos de capacitación determinado. Los núcleos son los únicos parámetros que se aprenden automáticamente durante el proceso de capacitación en la capa de convolución. Por otro lado, el tamaño de los granos, el número de granos, el relleno y la zancada son hiperparámetros que deben establecerse antes de que comience el proceso de capacitación (ver Tabla 3.1)

- Función de activación no lineal

  Las salidas de una operación lineal como la convolución se pasan a través de una función de activación no lineal. Aunque las funciones lisas no lineales, como

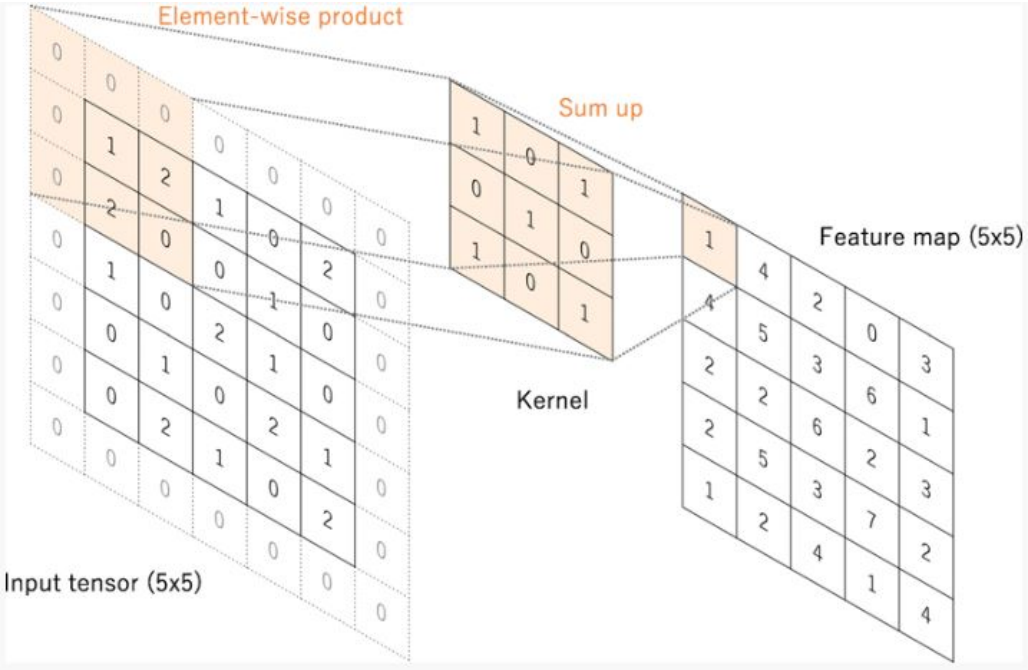

Figura 3.7: Una operación de convolución con relleno cero para retener las dimensiones en el plano [37].

la función de tangente sigmoidea o hiperbólica ($tanh$), se usaron anteriormente porque son representaciones matemáticas de un comportamiento biológico neuronal, la función de activación no lineal más común utilizada actualmente es la unidad lineal rectificada ($ReLU$), que simplemente calcula la función f de la Ecuación (3.5), estas funciones se muestran en la Figura 3.8.

$$f(x) = max(o, x) \tag{3.5}$$

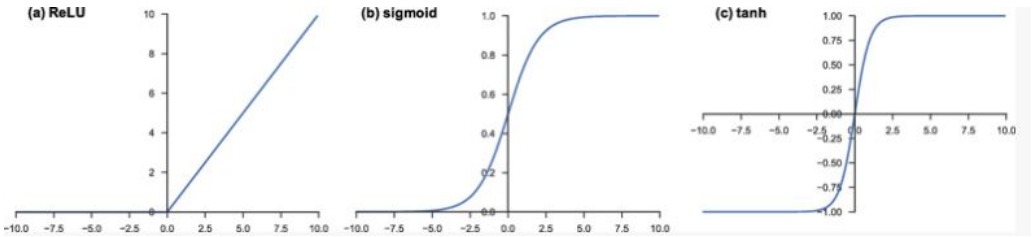

Figura 3.8: Funciones de activación comúnmente aplicadas a redes neuronales: una unidad lineal rectificada ($ReLU$), b sigmoidea y c tangente hiperbólica ($tanh$) [37].

- Capa de agrupación

  Una capa de agrupación proporciona una operación de disminución de muestreo típica que reduce la dimensionalidad en el plano de los mapas de características para introducir una invariancia de traducción a pequeños cambios y distorsiones, y disminuir el número de parámetros aprendibles posteriores. Es de notar que no hay ningún parámetro que se pueda aprender en ninguna de las capas de agrupación, mientras que el tamaño del filtro, la zancada y el relleno son hiperparámetros en las operaciones de agrupación, similares a las operaciones de convolución.

- Agrupación máxima

  La forma más popular de operación de agrupación es la agrupación máxima, que extrae parches de los mapas de características de entrada, genera el valor máximo en cada parche y descarta todos los demás valores (Figura 3.9). Una agrupación máxima con un filtro de tamaño 2×2 con un paso de 2 se usa comúnmente en la práctica. Esto reduce la dimensión en el plano de los mapas de entidades por un factor de 2. A diferencia de la altura y el ancho, la dimensión de profundidad de los mapas de entidades permanece sin cambios.

- Agrupación promedio global

  Otra operación de agrupación digna de mención es una agrupación promedio global. Una agrupación promedio global realiza un tipo extremo de disminución de resolución, donde un mapa de características con tamaño de: altura x ancho se muestrea en una matriz 1×1 simplemente tomando el promedio de todos los elementos en cada mapa de características, mientras que la profundidad de los mapas de características es retenido. Esta operación generalmente se aplica solo una vez antes de las capas completamente conectadas. Las ventajas de aplicar la agrupación promedio global son las siguientes: (1) reduce el número de parámetros que se pueden aprender y (2) permite que la CNN acepte entradas de tamaño variable.

- Capa completamente conectada

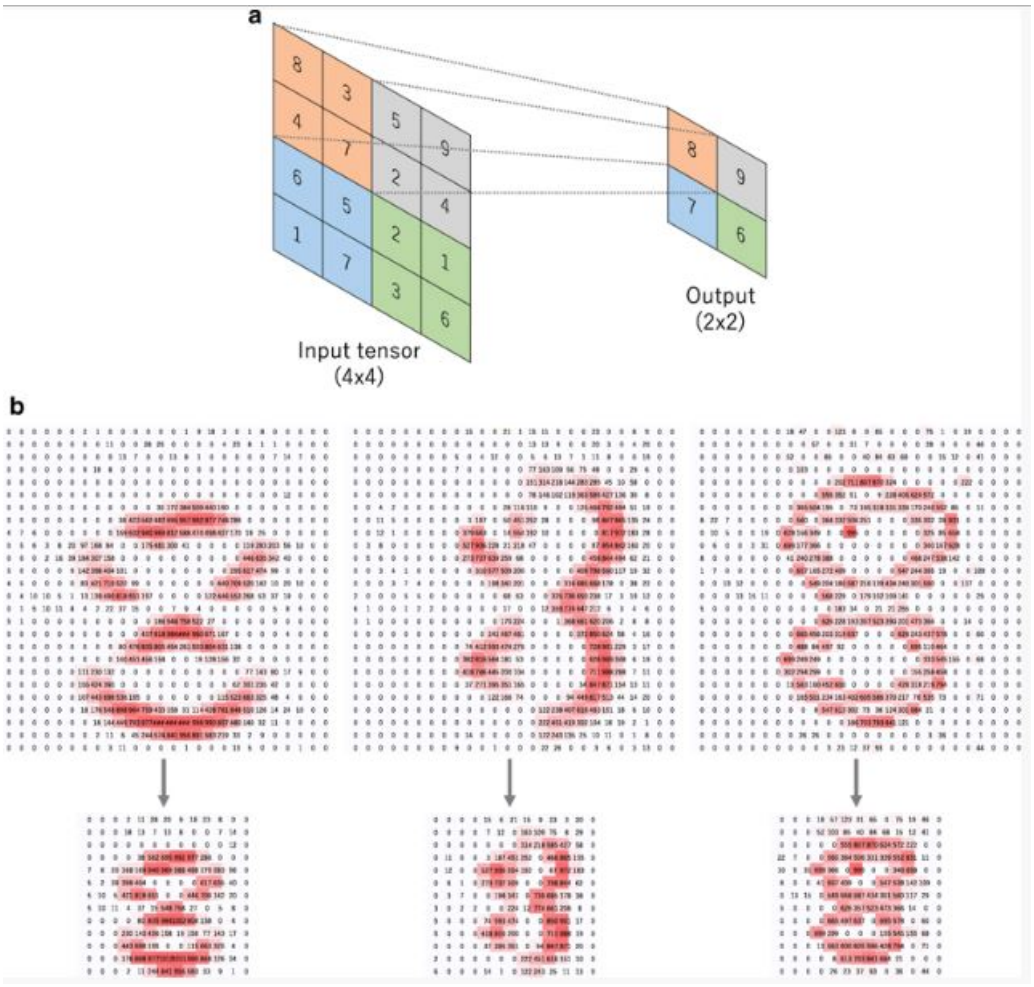

Figura 3.9: Un ejemplo de operación de agrupación máxima con un tamaño de filtro de 2×2 [37].

Los mapas de características de salida de la convolución final o capa de agrupación generalmente se aplanan, es decir, se transforman en una matriz unidimensional (1D) de números (o vectores) y se conectan a una o más capas completamente conectadas, también conocidas como capas densas, en el que cada entrada está conectada a cada salida por un peso que se puede aprender. Una vez que se crean las características extraídas por las capas de convolución y disminuidas por las capas de agrupación, se asignan mediante un subconjunto de capas completamente conectadas a las salidas finales de la red, como las probabilidades de cada clase en las tareas de clasificación. La capa final

totalmente conectada generalmente tiene el mismo número de nodos de salida que el número de clases. Cada capa completamente conectada es seguida por una función no lineal, como ReLU, como se describió anteriormente.

- Función de activación de la última capa

  La función de activación aplicada a la última capa totalmente conectada suele ser diferente de las demás. Se debe sear una función de activación apropiada de acuerdo con cada tarea. Una función de activación aplicada a la tarea de clasificación multiclase es una función $softmax$, de la Ecuación (3.6)

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_j}} \tag{3.6}$$

  donde z es un vector de las entradas a la capa de salida (si tiene 10 unidades de salida, entonces hay 10 elementos en z). Y nuevamente, j indexa las unidades de salida, entonces $j = 1, 2, ..., K$. La cual normaliza los valores reales de salida de la última capa completamente conectada a las probabilidades de la clase objetivo, donde cada valor oscila entre 0 y 1 y todos los valores suman 1. La función de activación para varios tipos de tareas se resume en la Tabla 3.2.

Tabla 3.2: Una lista de las funciones de activación de la última capa comúnmente aplicadas para diversas tareas [37].

Tarea	Función de activación de la última capa
Clasificación binaria	Sigmoideo
Clasificación de clase única multiclase	Softmax
Clasificación multiclase multiclase	Sigmoideo
Regresión a valores continuos	Identidad

- Técnica de entrenamiento *dropout*

  Para solucionar el problema de sobreajuste (*overfitting*) [21] que se produce cuando logra un buen ajuste de su modelo en los datos de entrenamiento, mientras que no se generaliza bien en datos nuevos e invisibles, o almenos para disminuirlo un poco. Se utiliza la técnica de entrenamiento *dropout*. La forma

en que lo hace, en la cual el *dropout* omite cada una de las neuronas de una forma aleatoría con una probabilidad p. En analisis se observó un pequeño aumento en el rendimiento en las CNNs entrenadas con *dropout*, en este caso variando las tasas de *dropout* en cada capa en un tamaño de 0.25, que dio los mejores resultados en la optimización

3.5.2. Redes neuronales recurrente

Las redes neuronales recurrentes (RNN) en [20] y [11], contienen bucles de retroalimentación, permitiendo que la información persista. Este tipo de redes procesan secuencia de datos, estos datos pueden ser texto, genomas, escritura, palabra hablada o series temporales numéricas, además de trabajar con imágenes.

En la siguiente Figura 3.10 se puede observar una parte de la red neuronal A, se observa alguna entrada x_t y genera un valor h_t. Un bucle permite que la información pase de un paso de la red al siguiente. La forma de bucle de la red neuronal es total-

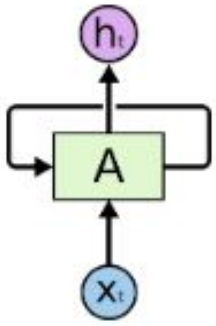

Figura 3.10: Red neuronal recurrente sin desenvolver [20].

mente equivalente a tener distintas redes neuronales, una para cada paso temporal t donde se permite que la información del instante $t-1$ sea transmitida a la red neuronal que computa el instante t. En la Figura 3.11 se muestra una RNN sin bucle. Uno de los atractivos de los RNN es la idea de que podrían conectar información previa a la tarea actual, como el uso de fotogramas de video anteriores para informar la comprensión del presente cuadro. Es por ello que una de las redes neuronales recurrentes más populares; es la *Long Short-Term Memory* (LSTM, red de memoria a

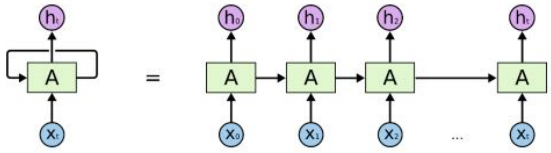

Figura 3.11: Red neuronal recurrente expandida [20].

corto y largo plazo), porque es capaz de asimilar dependencias *Long-Term* que redes neuronales recurrentes de otros tipos no son capaces de asimilar.

Para entender que es una dependencia *Long-Term* primero veamos que es una dependencia *Short-Term*(ver Figura 3.12). Consideremos que tenemos un modelo que intenta predecir cuál va a ser la siguiente palabra de una frase y estamos intentando predecir la palabra francés de la frase "En Francia la lengua oficial es el francés". En este caso, no se necesita mucho contexto para saber que la palabra va a ser francés, se dice entonces que el *gap* entre la información relevante (Francia, lengua oficial) y el lugar donde se necesita es pequeño. Una red neuronal recurrente cualquiera es capaz de aprender estas relaciones donde el *gap* es pequeño utilizando la información pasada. Veamos ahora el caso donde la dependencia es *Long-Term* (ver Figura 3.13),

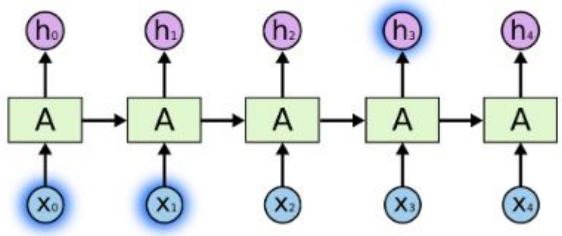

Figura 3.12: Ejemplo de gap pequeño [11].

es decir, un caso en que el gap entre la información relevante y el lugar donde se necesita es grande. Si de nuevo tenemos el modelo que predice palabras, pero ahora el texto es: "Crecí en Francia... Hablo francés", donde los puntos suspensivos reflejan otro texto entremedio, se puede ver como la distancia (*gap*) es mayor. Cuanto mayor es esta distancia, más dificultades tienen las redes neuronales recurrentes para conectar la información. Aun así, en teoría las redes neuronales recurrentes son capaces de aprender estas dependencias *Long-Term*, aunque a veces no sea así. Ante este problema, aparecieron las LSTM, que no tienen este problema ya que están

diseñadas específicamente para poder aprender de estas relaciones *Long-Term*.

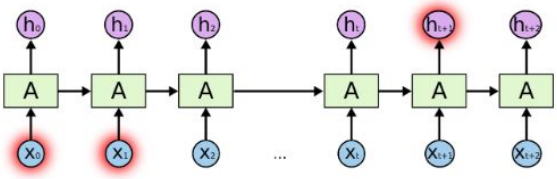

Figura 3.13: Ejemplo de gap grande [11].

3.5.3. Red de memoria a corto y largo plazo

Las redes de memoria a corto y largo plazo, generalmente llamadas "LSTM" [11], son un tipo especial de RNN, capaces de aprender dependencias a largo plazo. Fueron introducidos por Hochreiter y Schmidhuber (1997). Las LSTM están diseñadas explícitamente para evitar el problema de dependencia a largo plazo. Recordar información durante largos períodos de tiempo es prácticamente su comportamiento predeterminado.

Para entender bien el funcionamiento de las LSTM, es importante analizar su estructura, en la Figura 3.14 se representa la forma de cadena de una LSTM y los vectores y operaciones que contiene, donde cada línea representa un vector que conecta el output del instante $t-1$ con el input del instante t, cada circulo rosa representa una operación entre vectores y cada caja amarilla una capa de la red neuronal donde la información pasa por una función. Hay dos componentes de gran

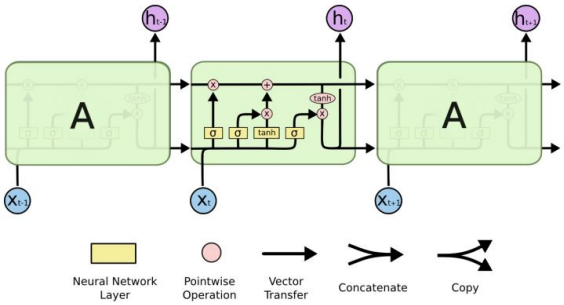

Figura 3.14: Módulo de una LSTM con sus cuatro capas [11].

importancia dentro de la estructura de la LSTM, estas son el *cell state* y las *gates*.

El *cell state* es el vector que recorre la parte superior de la LSTM en la 3.15 y su importancia viene dada porque transporta la mayoría de la información entre iteraciones. La información del *cell state* es alterada únicamente por interacciones lineales, por lo que al no sufrir ninguna transformación mayor la información nunca es alterada de gran manera en una sola iteración, permitiéndole recordar información del pasado con más facilidad.

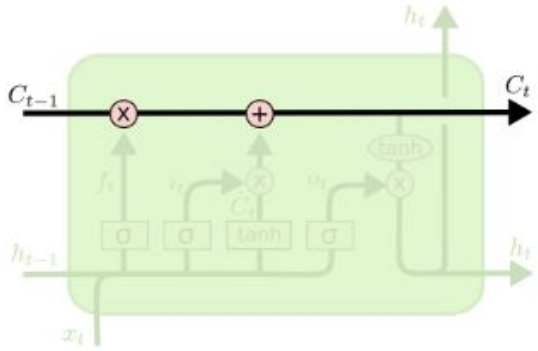

Figura 3.15: *Cell state* de una LSTM [11].

El otro componente de vital importancia en el funcionamiento de la LSTM son las *gates* (puertas) como se muestra en la Figura 3.16. Las *gates* están compuestas de una función (normalmente una sigmoide) y una operación puntual entre vectores. Como bien dice su nombre, funcionan como puertas de entrada para la información, donde la información de un vector pasaría por la capa sigmoidea produciendo un número entre 0 y 1, siendo 0 igual a "toda esta información es irrelevante" o 1 "toda esta información es relevante". En función de la importancia que esta capa sigmoidea da a la información del vector, esta pasa después al *cell state* a través de una operación entre vectores.

Hay distintos tipos de *gate* en una LSTM, el primer tipo es la llamada *forget gate layer*. Como bien indica su nombre, esta es la encargada de decidir qué información se olvida de la iteración anterior, es decir, se trata de una capa sigmoidea que en la iteración t, tiene en cuenta la información h_{t-1} (que se corresponde al output de la iteración anterior) y x_t (*input* de la presente iteración t) y decide qué información es relevante. Eso lo hace pasando esta información por una sigmoide que genera un valor entre 0 y 1 para cada elemento del vector de información del *cell state* C_{t-1}.

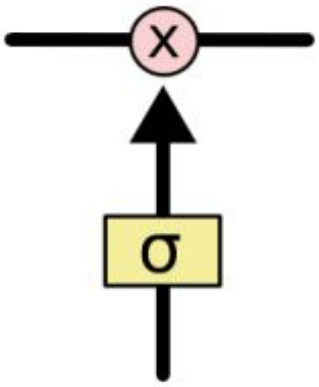

Figura 3.16: Representación de una *gate* [11].

Como se puede ver en el diagrama inferior, el vector que sale de la sigmoide pasa entonces a multiplicar el *cell state*, haciendo que olvide aquella información que se ha multiplicado por un 0 y que recuerde la otra en menor o mayor medida. En Figura 3.17 se puede observar también la función de la *forget layer*. Dónde σ representa la función sigmoidea, W_f son los *weights* (pesos) y b_f el sesgo.

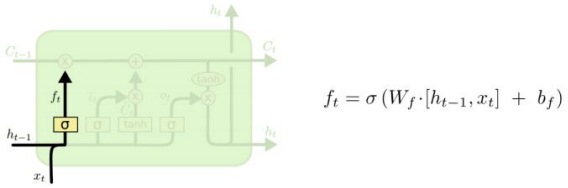

Figura 3.17: Representación de la *forget gate layer* [11].

Otros dos tipos de las *gates* que hay dentro de una LSTM son la *input gate layer* y la *tanh layer* como se muestra en la Figura 3.18. Estas dos son los componentes de la LSTM encargados de decidir qué información nueva es guardada en el *cell state*. La *input gate layer* (i_t en el diagrama) es una capa sigmoidea encargada de decidir que valores van a ser actualizados, mientras que la capa tanh (una tangente hiperbòlica) crea un vector $\tilde{C}_t$ que contiene los candidatos a ser añadidos al *cell state*.

Con la información resultante i_t y $\tilde{C}_t$ se actualiza el *cell state* de la iteración anterior al de esta. Recordemos que hasta ahora al *cell state* únicamente se le había aplicado el producto resultado de la *forget gate layer*, por lo que seguía teniendo únicamente la información de $t-1$ habiendo olvidado aquella que era considerada

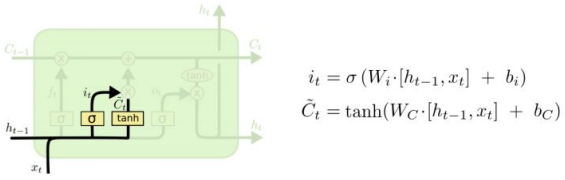

Figura 3.18: *Input gate layer* y *tanh layer* en una LSTM [11].

irrelevante. A este *cell state* C_{t-1} se le suma ahora el producto $i_t * \tilde{C}_t$ como puede verse a continuación en la Figura 3.19. Tras esta transformación, ya se tiene el *cell state* correspondiente a la iteración t, C_t, que es el que se transmitirá a la siguiente iteración.

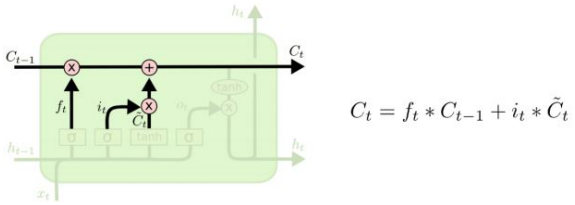

Figura 3.19: Actualización del *cell state* [11].

Para acabar, solo falta saber cómo se obtiene el output de esta iteración, es decir, h_t. Éste es el resultado de parte de la información deseada del *cell state* y parte de la información resultante del *input* de la iteración presente y el output de la anterior. Como se puede ver en la Figura 3.20, la información del *cell state* pasa por una hipertangente transformando todos los valores de este en valores entre -1 y 1, y lo multiplica por el resultado de la sigmoide con información del output pasado y el *input* presente. Esta información pasará a ser el output de la iteración t, que será luego usado como *input* en la iteración $t + 1$.

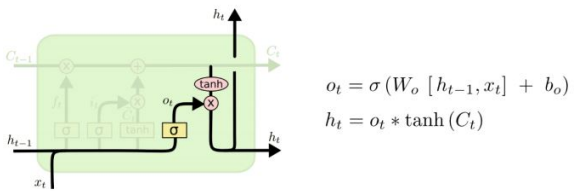

Figura 3.20: Cálculo del *output* de una LSTM [11].

3.5.4. Optimización

Al trabajar con redes neuronales [23] lo que realmente nos interesa optimizar durante el entrenamiento es alguna medida P que representa el rendimiento de la red. El inconveniente que surge es que P suele ser intratable (No hay un algoritmo que dé una solución a algún problema en tiempo polinomial.) y depende del conjunto de prueba, no del conjunto de entrenamiento. En consecuencia, lo que se hace es optimizar P de manera indirecta, recurriendo a otra función de costo L, con la esperanza de que esto también mejore P.

El objetivo de un algoritmo de aprendizaje es reducir el error de generalización, que se expresa en la Ecuación 3.7

$$R(\theta) = E_p[L(F(x, \theta), y)], \tag{3.7}$$

donde p es la distribución de los pares (x, y), L es la función que cuantifica la pérdida para cada par y $f(x, \theta)$ es la salida de la red. Esta cantidad no es más que la pérdida en la que se incurre para un determinado valor de los parámetros θ y es por esto que también se suele llamar riesgo. Generalmente no conocemos p, sino que tenemos un conjunto de entrenamiento, por lo que minimizamos el costo sobre este conjunto en vez de sobre toda la distribución, con lo cual ahora estamos minimizando el riesgo empírico (ver Ecuación 3.8) y esperamos que esto a su vez nos permita disminuir $R(\theta)$ de manera significativa. Sin embargo, este enfoque es propenso al *overfitting*, ya que modelos con alta capacidad pueden simplemente memorizar el conjunto de entrenamiento. Además, muchas técnicas efectivas de optimización se basan en descenso por gradiente, mientras que funciones como la que usamos para cuantificar la exactitud de la red no tienen derivadas útiles. Debido a esto es raro que se recurra a la minimización directa del riesgo empírico en redes neuronales, sino que se utilizan funciones substituto para la pérdida, como pueden ser el costo cuadrático, la entropía cruzada, o el negativo *log-likelihood*.

$$J(\theta) = E_{\hat{p}}[L(F(x, \theta), y)] = \frac{1}{n} \sum_{k=1}^{n} [L(F(x_k, \theta), y_k), \tag{3.8}$$

- Adam

Adam (Adaptive Moment Estimation) es otro método que calcula tasas de aprendizaje que se adaptan para cada parámetro. En adición a guardar una media móvil exponencial de los gradientes al cuadrado, este algoritmo también guarda una media móvil de los gradientes (sin el cuadrado). Los promedios de los gradientes y sus cuadrados tienen entonces las fórmulas tales como en las Ecuaciones 3.9 y 3.10.

$$m_t = \alpha m_{t-1} + (1 - \alpha) g_t, \tag{3.9}$$

$$v_t = \beta v_{t-1} + (1 - \beta) g_t^2. \tag{3.10}$$

Se puede pensar m_t como un estimador del primer momento de los gradientes, mientras que v_t es un estimador del segundo momento; de ahí viene el nombre del método. Dado que m_t y v_t son inicializados como vectores cero, esto ocasiona que estén sesgados hacia el cero, particularmente durante los pasos iniciales, cuando las tasas de decaimiento son pequeñas. Para contrarrestar esto, se introducen unos estimadores corregidos (ver Ecuación 3.11):

$$\hat{m}_t = \frac{m_t}{1 - \alpha^t}, \hat{v}_t = \frac{v_t}{1 - \beta^t}. \tag{3.11}$$

Estos son los que luego se usan para actualizar los parámetros (ver Ecuación 3.12):

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \tag{3.12}$$

Los valores recomendados para β, α y ϵ suelen ser 0,9, 0,999 y 10^{-8} respectivamente.

Capítulo 4

Diseño

El desarrollo de este trabajo se basa en las fases de ciencia de datos que hacen que el desarrollo de cualquier proyecto sea más fácil, más preciso o acertado. En el desarrollo de cualquier proyecto se inicia con la fase de la definición del proyecto, en donde se establecen los objetivos de este mismo y la fase final es la ejecución de acciones (ver sección 3.3). Basándose en estas fases se propone un algoritmo para la clasificación automática de eventos delictivos en noticias periodísticas, mediante la implementación de una red neuronal convolucional y la red LTSM. Posteriormente, con los resultados clasificados se realiza el análisis de las noticias, con la finalidad de detectar el estado con mayor índice de delitos en el país.

En este capítulo se explican las fases que corresponden a la clasificación y posteriormente la fase que corresponde al análisis de las noticias clasificadas.

4.1. Diseño propuesto para la clasificación de eventos

A continuación se describe el diseño de la propuesta de solución para la clasificación de eventos delictivos en noticias periodísticas. En la cual se incluye el desarrollo de una red neuronal convolucional y la red LTSM basada en aprendizaje profundo. En la Figura 4.1 se puede observar las etapas que se deben seguir para llevar acabo está fase.

El algoritmo propuesto consiste de los siguientes pasos:

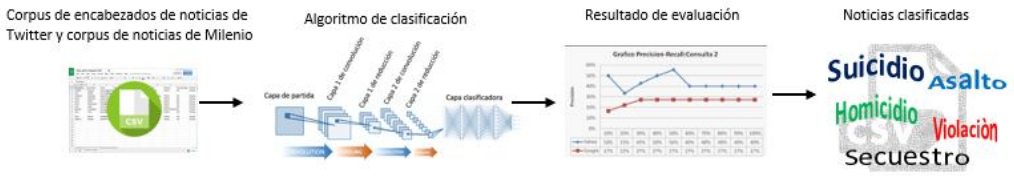

Figura 4.1: Diseño propuesto para la clasificación automática de eventos delictivos.

1. Entrada: corpus de entrenamiento (encabezado de noticias de Twitter) y corpus de prueba (noticias de la página local Milenio).

2. Aplicación del algoritmo de clasificación a través de redes neuronales (CNN y LTSM).

3. Evaluación del modelo obtenido con los datos de prueba.

4. Salida: noticias clasificadas.

A continuación, se describe de manera más general cada uno de estos pasos.

4.1.1. Corpus de entrenamiento

El corpus para entrenamiento se recuperó de [24] el cual cuenta con encabezados de noticias obtenido de la red social de Twitter, en este artículo se mencionan seis clases de eventos delictivos los cuales son: homicidio, violación, asalto, suicidio, secuestro y explotación sexual. En la Figura 4.2 se muestran los datos con lo que se cuenta para el entrenamiento. Estos datos incluyen el número del tweet, fecha de la publicación del tweet, medio en el que se transmitió, el tweet que contiene el encabezado de la noticia y la etiqueta (clase) a la que pertenece esa noticia.

numero	fecha	medio	tweet	etiqueta
9.32E+17	Fri Nov 17 22:38:04 CST 2017	NTelevisa_com	#EnPunto Peligra atención médica en zona serrana de #Chihuahua tras el secuestro del médico Blas Godínez& https://t.co/j4	Secuestro
9.32E+17	Fri Nov 17 23:07:04 CST 2017	NTelevisa_com	Se metieron a robar un cajero automático con todo y camionetas a un centro comercial de Ecatepec en el #Edomex& https://	Asalto
9.32E+17	Sat Nov 18 00:27:23 CST 2017	NTelevisa_com	Con camionetas y cuerdas intentaron robar un cajero del centro comercial Plaza Aragón, en Ecatepec, #Edomex.& https://t.c	Asalto
9.32E+17	Sat Nov 18 12:01:01 CST 2017	SSP_CDMX	Detuvimos a tres sujetos acusados de robar 24 mil pesos a un hombre de 28 años que ingresaba a su casa en la col. I& https://	Asalto
9.32E+17	Sat Nov 18 14:07:34 CST 2017	elsolde_mexico	Comando ejecuta a dos personas en un bar en Nuevo León https://t.co/Cw9b7mQWJk https://t.co/k6QWyYurBg	Homicidio
9.32E+17	Sat Nov 18 17:00:01 CST 2017	lajornadaonline	#Morrissey cuestiona ola de denuncias de acoso sexual en #Hollywood https://t.co/vXG0zNgnS1 https://t.co/O88ioglbNK	Violación
9.32E+17	Sat Nov 18 18:16:00 CST 2017	Milenio	El #Vaticano investigará abusos sexuales en escuela de #Roma después de la publicación del libro 'El pecado origina& https:/	Violación
9.32E+17	Sat Nov 18 18:40:00 CST 2017	Pajaropolitico	La violencia aumentó en el país aún con la gendarmería: entre 2015 y 2017 los homicidios dolosos se han incrementad& https	Homicidio
9.32E+17	Sat Nov 18 19:36:00 CST 2017	Milenio	'Naturista' envenena a 14 y mata a otra en #Oaxaca https://t.co/GW8l3R9xbG https://t.co/JO7DfLe2pb	Homicidio
9.32E+17	Sat Nov 18 20:10:00 CST 2017	El_Universal_Mx	Los homicidios ocurrieron en los municipios de Naucalpan, Ecatepec y Chalco https://t.co/oPaZSNf8uy	Homicidio
9.32E+17	Sat Nov 18 20:34:01 CST 2017	Pajaropolitico	A un año de que @EPN termine, @anairamzepol y @ccastan3da, reflexionan sobre el aumento de homicidios en México.& h	Homicidio
9.32E+17	Sun Nov 19 00:17:00 CST 2017	Pajaropolitico	Los niños indígenas en México no tienen muchas opciones: son reclutados por el narco o víctimas de homicidio o desa& http	Homicidio
9.32E+17	Sun Nov 19 04:30:00 CST 2017	NoticiasMVS	.@Uber encara demanda colectiva en #EEUU por agresiones sexuales de conductores https://t.co/zIIb5Yk21p	Violación
9.32E+17	Sun Nov 19 11:55:25 CST 2017	Notimex	15 personas murieron y 5 resultaron heridas durante una avalancha humana en #Marruecos, mientras se repartía ayuda& htt	Homicidio
9.32E+17	Sun Nov 19 13:25:00 CST 2017	El_Universal_Mx	La PGJ cumplimentó las órdenes de aprehensión en su contra por los delitos de robo a casa habitación y robo de vehí& https:	Asalto
9.32E+17	Sun Nov 19 13:45:00 CST 2017	El_Universal_Mx	El asesinato fue perpetrado por hombres armados en el llamado Triángulo Rojo de robo de hidrocarburo https://t.co/aEkdYb	Asalto
9.32E+17	Sun Nov 19 15:55:02 CST 2017	GoogleNewsMX	Silencioso aumento: precio del gas se disparó en más de 30% durante el año https://t.co/hkT3pdT2uL	Suicidio
9.32E+17	Sun Nov 19 16:49:08 CST 2017	Reforma	El director de Izzi, Adolfo Lagos, fue baleado por asaltantes que le robaron una bici en Teotihuacán https://t.co/GnR44RRByD	Homicidio
9.32E+17	Sun Nov 19 18:26:38 CST 2017	NTelevisa_com	Los homicidios contra mujeres y los #feminicidios se duplican en #Zacatecas; en lo que va del año se han registrado& https:/	Homicidio
9.32E+17	Sun Nov 19 19:12:30 CST 2017	Milenio	#PGR colaborará en caso de asesinato de directivo de #Izzi: @EPN https://t.co/wIdLZ2wo28 https://t.co/SNFuQoEMfE	Homicidio
9.32E+17	Sun Nov 19 19:30:01 CST 2017	elsolde_mexico	#Almomento | @EPN condena asesinato del vicepresidente de @Televisa; @PGR_mx cooperará en investigación& https://t	Homicidio
9.32E+17	Sun Nov 19 19:51:32 CST 2017	Siete24Noticias	La @PGR_mx participará en la invetigación del homicidio del vicepresidente de @Televisa https://t.co/SVcwHTHZ8f	Homicidio
9.32E+17	Sun Nov 19 20:45:29 CST 2017	El_Universal_Mx	Después de intentar robar a una familia, los delincuentes se escondieron en en una vecindad, lo que provocó un inte& https	Asalto
9.32E+17	Sun Nov 19 22:12:00 CST 2017	Milenio	#PGR colaborará en caso de asesinato de directivo de #Izzi: @EPN https://t.co/wIdLZ2eMDy https://t.co/qogt5Xhve3	Homicidio
9.32E+17	Sun Nov 19 22:40:01 CST 2017	NTelevisa_com	Los homicidios contra mujeres y los #feminicidios se duplican en #Zacatecas; en lo que va del año se han registrado& https:/	Homicidio
9.32E+17	Mon Nov 20 00:03:56 CST 2017	Reforma	Charles Manson, quien ordenó una serie de asesinatos en 1969, murió hoy a los 83 años https://t.co/oEBdHNuDF5	Homicidio

Figura 4.2: Encabezado de noticias de twitter [24].

En este caso para llevar a cabo la clasificación de nuestro interés se tomaron cinco clases, las cuales son: homicidio, secuestro, asalto, suicidio y violación.

4.1.2. Corpus de prueba

En el caso del corpus para probar, la fuente de información por la que se optó es Milenio una página de noticias local, como se mencionó anteriormente, debido a que esta página nos proporciona un apartado de noticias Policíacas en la cual contiene las noticias de eventos delictivos, además de que cuenta con noticias de toda la república mexicana. También se revisaron noticias de otros periódicos digitales, tales como: el Sol de Puebla, el Sol de México y la Prensa, pero sólo se proporciona información de temas en general, por lo que no fueron considerados.

Con ayuda del lenguaje de programación Python, se crea un raspado web con las bibliotecas scrapy [27] y BeautifulSoup [5]. En el Algoritmo 1 se muestran los pasos que se siguieron para la obtención de las noticias periodísticas por medio de la página local: Milenio. El algoritmo consiste en la identificación de los elementos más importantes de la página local milenio los cuales son: título, contenido, fecha y lugar de la noticia, posteriormente se recupera la información de estos elementos en modo texto, donde se garantiza que dicha información no contenga basura en este caso anuncios de algún otro tema o de alguna otra noticia, finalmente se estructura y se genera un archivo csv con la información recuperada de la noticia. En la Figura 4.3 se

muestran los elementos principales extraídos, con ayuda de la función *add_xpath()* se recupera el texto de las etiquetas título, fecha y lugar y para la etiqueta de contenido se recupera el texto con la función *soup.find* de la página html.

Algoritmo 1 Algoritmo de raspado web.

Entrada: Página local Milenio

Salida: Archivo estructurado csv que contiene título, contenido, fecha y lugar de la noticia

1: Definir las etiquetas a extraer de la página HTML:
2: Título
3: Contenido
4: Fecha
5: Lugar
6: Obtener la URL de la sección "Policía" de la página local (Milenio)
7: Obtener el xpath de la etiqueta título, fecha y lugar de la página HTML
8: Obtener el identificador de la etiqueta contenido de la página HTML
9: Ejecución del código para la recuperación del texto con las funciones *add_xpath* y *soup.find*, de las etiquetas título, contenido, fecha y lugar
10: Eliminar noticias duplicadas.

Figura 4.3: Variables extraídas de la página local (Milenio).

4.1.3. Etiquetado de noticias

Al obtener el corpus de noticias, se procede a realizar el etiquetado manual de las noticias, este etiquetado consiste en la lectura de la noticia y posteriormente al haber comprendido el tema de está, se etiqueta. Las clases considerados son los siguientes: homicidio, secuestro, asalto, suicidio y violación. En la Figura 4.4 se muestra un ejemplo de este etiquetado.

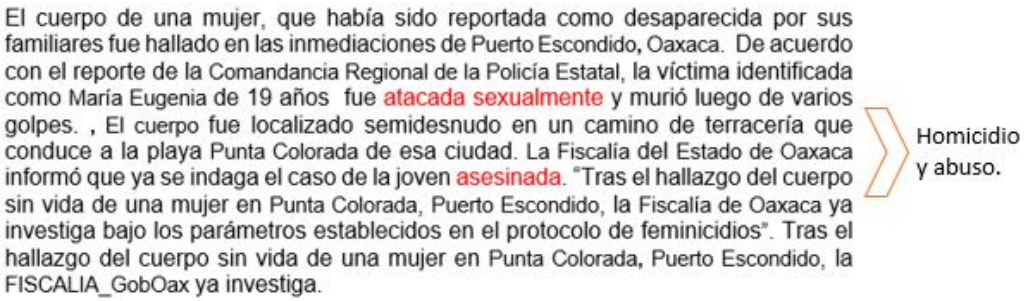

Figura 4.4: Ejemplo de etiquetado del corpus de noticias.

4.1.4. Aplicación del algoritmo de clasificación

Para llevar a cabo la primera fase, se realizara una prueba de red neuronal convolucional (CNN) y una red neuronal llamada: memoria a corto y largo plazo (LSTM), en donde las funciones de activación [37] que dieron buenos resultados son las de *relu*: esta función de activación es una de las más utilizadas y la función de activación *softmax* especialmente para la clasificación multiclase. En el caso de la optimización se realizaron pruebas con los optimizadores *adam* , *nadam*, *adadelta*, *adagrad* y *SGD*, en donde el optimizador que mejores resultados arrojo es el de *adam*. Para la técnica de *dropout* los mejores resultados se obtuvieron con la medida de 0.25, además de que se encontró un estudio [21] donde esta medida dio buenos resultados. Finalmente, se eligio la métrica de *Accuracy* y la aplicación de las 10 epocas, pues se obtuvieron buenos resultados al realizar las pruebas de clasificación.

Los pasos que se siguieron se muestran en el Algoritmo 2:

Algoritmo 2 Algoritmo de clasificación basado en redes neuronales.

Entrada: Corpus para entrenar, corpus para realizar las pruebas y corpus de palabras preentrenadas [31].

Salida: Corpus de noticias clasificadas en un csv.

1: Usando los datos de entrenamiento se genera una matriz de incrustaciones de palabras, usando una representación para cada palabra. Cada palabra similar, tendrá una representación similar.
2: Aplicación de la red neuronal convolucional (Conv1D) utilizando la función de activación *relu* entre otras.
3: Aplicación de LSTM, donde se utilizó la función de activación *relu* y *softmax*, un Dropout(0.25) entre otras.
4: Se aplica el optimizador *Adam*.
5: Se aplica la métrica *Accuracy*.
6: Se realiza la prueba con 10 épocas.

4.2. Diseño propuesto para el análisis de la clasificación

Después de la obtención de las noticias se procede a realizar un análisis de las mismas. El análisis consta de las siguientes fases (ver Figura 4.5):

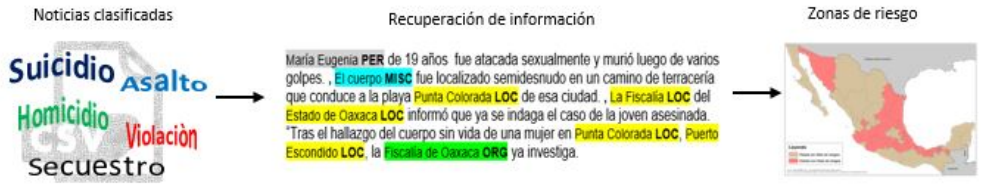

Figura 4.5: Análisis propuesto de las noticias etiquetadas.

1. Datos de entrada: noticias clasificadas.
2. Aplicación del algoritmo de reconocimiento de entidades con nombre.
3. Datos de salida: Zonas de riesgo.

A continuación, se describe a profundidad el procedimiento para el análisis de las noticias.

4.2.1. Aplicación del algoritmo de reconocimiento de entidades con nombre

Se procede a realizar la recuperación de entidades, en este caso se aplica el reconocimiento de entidades con nombre, descrita en la sección 3.2.2. Esta tarea se realizó con ayuda de la librería de Python spacy [29], donde se utilizó el modelo *es_core_news_sm* que proporciona spacy para el idioma español, el cual asigna vectores de token específicos de contexto, etiquetas POS, análisis de dependencia y entidades con nombre. Admite la identificación de entidades como nombres de personas (PER), organizaciones (ORG), la ubicación definida política o geográficamente (LOC) y entidades diversas (MISC). En la Figura 3.3 se presenta un ejemplo del resultado de la aplicación del algoritmo de reconocimiento de entidades con nombre

a una noticia de la página local Milenio. Como puede observarse se identificaron estados tales como el de Oaxaca (LOC), organizaciones como la fiscalía de Oaxaca (ORG), personas como María Eugenia (PER) y entidades diversas tales como: El cuerpo (MISC).

En este caso para la obtención de los estados de alto riesgo en la república mexicana, en el Algoritmo 3 se muestran los pasos que se siguieron para la obtención de estos. Como datos de entrada se requieren las noticias clasificadas obtenidas de la primera fase y un diccionario creado de estados y municipios obtenidos de [2]. Posteriormente se aplica el algoritmo NER a cada noticia obteniendo la ubicación definida política y geográficamente (LOC), una vez obtenidas estas etiquetas de las noticias se procede a ubicarlas en el diccionario de estados y municipios, finalmente se realiza un recuento para la obtención de los estados con mayor mención en cada noticia y se realiza la graficación de estos resultados con ayuda de la biblioteca Matplotlib [12] que proporciona Python. Como resultados finales se obtienen los estados con mayor incidencia de estos actos delictivos.

Algoritmo 3 Algoritmo de estados con mayor incidencia

Entrada: Corpus de noticias clasificadas, diccionario de estados y municipios.

Salida: Estados con mayor incidencia.

```
1: estados := []
2: para noticia en corpus hacer
3:     ubicacionesLOC := algoritmoNER(noticia)
4:     estadosMunicipios := busquedaDiccionario(ubicacionesLOC, diccionario)
5:     estadosnoticia := busquedaMayorMencion(estadosMunicipios)
6:     estados := estados + estadosnoticia
7: fin para
8: graficacion(estados)
```

A continuación en el siguiente capítulo se presentan los resultados experimentales obtenidos hasta el momento.

Capítulo 5

Resultados

En este capítulo se muestran los resultados experimentales de la clasificación de eventos delictivos en noticias periodísticas, así mismo se muestran los resultados que se obtienen después del análisis de las noticias clasificadas. También se muestra el conjunto de datos con la información que se requirió para el entrenamiento y la prueba en la ejecución de las redes neuronales. Posteriormente se muestra la exactitud, exhaustividad, precisión y la métrica F_1 global de esta clasificación, así mismo la exactitud obtenida por cada evento delictivo. Finalmente se muestran los resultados obtenidos en el análisis de las noticias clasificadas, en este análisis se realiza una visualización de las palabras más representativas de cada evento delictivo, además se muestran las zonas de riesgo en homicidio, secuestro, asalto, suicidio y violación. Además de mostrar los grupos vulnerables de víctimas por edades y sexo, así como los culpables de estos delitos.

5.1. Conjunto de datos

Para el conjunto de datos se cuenta con dos corpus, uno para entrenamiento y otro para prueba. En este caso los datos para realizar el entrenamiento se muestran en la Tabla 5.1 recuperados de [24], este contiene: los temas de encabezados de noticias de Twitter con las etiqetas respectivas de los eventos delictivos que se clasificarán, la cantidad de noticias y su respectivo vocabulario por tema, además se realiza un recuento total de noticias y su vocabulario.

Tabla 5.1: Datos para entrenar.

Tipo de noticias	Cantidad de noticias	Total de vocabulario
Homicidio	4,352	95,734
Suicidio	334	5,844
Asalto	2,997	59,568
Secuestro	365	7,183
Violación	666	13,704
Total	8,714	182,033

El segundo corpus es usado para realizar las pruebas correspondientes. Como se muestra en la Tabla 5.2, el total de noticias anotadas es de 577 (Gold) recuperadas del 06 de agosto del 2019 al 23 de marzo del 2020. El total de noticias identificadas como homicidio es de 367, en suicidio el total de noticias identificadas son 16, en asalto se identificaron 65 noticias, se identificaron 91 noticias en el evento criminal de secuestro y para violación se identificaron 38 noticias. La columna tres muestra el total de vocabulario por clase, además de mostrar el total de noticias y el total de vocabulario.

Tabla 5.2: Datos para probar.

Tipo de noticias	Gold	Total de vocabulario
Homicidio	367	144,870
Suicidio	16	4,637
Asalto	65	17,196
Secuestro	91	34,558
Violación	38	14,372
Total	577	215,633

5.2. Resultados experimentales de la clasificación

En los resultados experimentales se realizó la prueba con 577 noticias, donde se obtuvieron los siguientes resultados de clasificación (ver Tabla 5.3), se obtuvieron 366 noticias clasificadas en el sistemas de un total de 367 noticias etiquetadas en homicidio. En la clase de suicidio de un total de 16 noticias, el sistema clasificó 5 noticias en esta clase. En asalto el sistema arrojo 95 noticias de un total de 65 noticias etiquetadas. En secuestro 91 noticias etiquetadas el sistema clasificó 69 y en violación el sistema clasificó 42 noticias de un total de 38 noticias etiquetadas. Además se puede observar que en homicidio se clasificaron correctamente 315 noticias de las 367 etiquetadas. En suicidio se clasificaron correctamente 5 noticias, las noticias que arrojo el sistema son las mismas que se clasificaron. Para asalto se clasificaron correctamente 54 noticias. En secuestro se clasificaron correctamente 47 noticias y para violación se clasificaron correctamente 22 noticias.

Tabla 5.3: Resultados de la clasificación.

Tipo de noticias	Gold	Clasificadas	Correctas
Homicidio	367	366	315
Suicidio	16	5	5
Asalto	65	95	54
Secuestro	91	69	47
Violación	38	42	22

Finalmente, como se muestra en la Tabla 5.4 se analizó el Gold el cual cuenta con 577 noticias anotadas, este análisis consiste en la participación de tres expertos, cada uno realizo la lectura de estas noticias y respectivamente etiqueto la noticia acorde al tema que correspondía cada una de estas. Se optó por la elección de los resultados obtenidos con el U1 (Usuario 1), en donde se obtuvo una exactitud global de clasificación del 77 %, esto indica que la mayoría de los resultados se clasificaron correctamente. También se aplicó la métrica de precisión obteniendo 73 %, en ex-

haustividad se obtuvo un 62 % y en la métrica F_1 se obtuvo un 77 %. Se utilizaron las métricas que nos proporciona la biblioteca scikit-learn [1] de Python para obtener la evaluación de la predicción.

Tabla 5.4: Análisis del Gold.

Métrica	U1	U2	U3
Exactitud	77 %	75 %	76 %
Precisión	73 %	74 %	74 %
Exhaustividad	62 %	70 %	60 %
F_1	77 %	74 %	75 %

Analizando otros métodos de clasificación de eventos delictivos se observó, por ejemplo, que en Reyes-Ortiz y Bravo [24] se basan en patrones para reconocer y extraer eventos delictivos de noticias publicadas en periódicos digitales mexicanos, donde dichos patrones se mejoran con información morfológica y categorías POS. Los eventos delictivos considerados son: homicidio, violación, asalto, suicidio, secuestro y explotación sexual. En la evaluación se obtuvo una precisión del 0.719, en exhaustividad un 0.573 y en la métrica F_1 0.636, donde se demuestra una buena efectivadad general para reconocer eventos delictivos. Otro caso se presenta en Silva [28] profundiza sobre el nivel de criminalidad presente en las noticias policiales, distinguiendo siete temáticas las cuales son: delitos sexuales, incendios, drogas, disturbios, homicidios, tránsito y robos. Para la evaluación de los modelos de clasificación se obtuvieron los siguientes resultados, los modelos Naive Bayes mostraron un bajo desempeño para los esquemas de 700 y 1400 palabras con un accuracy de 71.12 % y 68.20 % respectivamente. Por otro lado, los modelos K-nn mostraron resultados bastantes similares entre los esquemas de 700 y 1.400 palabras, siendo levemente superior el esquema con 700 palabras, con un accuracy de 87.33 %, precisión 88.15 %, recall 85.74 % y F-measure 86.76 %. En comparación con estos métodos aplicados para la clasificación de diferentes delitos, en la aplicación de redes neuronales se obtiene una buena

medida para la clasificación, dado que no se realizó un análisis profundo para el texto de las noticias.

En la Tabla 5.5 se observan las métricas de exactitud, exhaustividad, precisión y F_1 obtenidas en la evaluación para homicidio, suicidio, asalto, secuestro y violación.

Tabla 5.5: Resultados de la evaluación de la clasificación.

Tipo de noticias	Exactitud	Exhaustividad	Precisión	F_1
Homicidio	86 %	86 %	99 %	92 %
Suicidio	31 %	31 %	99 %	48 %
Asalto	83 %	83 %	99 %	90 %
Secuestro	52 %	52 %	99 %	68 %
Violación	58 %	58 %	99 %	73 %

5.3. Resultados experimentales del análisis de las noticias

Como primer análisis en las noticias se realizó una nube de palabras para la visualización de estas, donde se eliminan las palabras vacías las cuales no tienen mucho significado. De estas palabras se indica su importancia y frecuencia por el tamaño de cada palabra con ayuda de la biblioteca Wordcloud [3] y para la visualización con la biblioteca Matplotlib [12] que proporciona Python. A continuación, se describe por clases (homicidio, secuestro, asalto, suicidio y violación) los resultados obtenidos.

En la Figura 5.1 correspondiente al evento delictivo de homicidio se observan como palabras más importantes las remarcadas en color rojo, entre estas palabras están: feminicidio, violencia, víctima, asesinato, delito, cuerpo, homicidio, niña, niño, mujer, etc. Estas palabras son distintivas en estas noticias para este evento delictivo.

Figura 5.1: Gráfica de palabras distintivas en la clase homicidio.

En el evento de asalto, las palabras más representativas a este evento delictivo son: asalto, víctima, robo, delito, arma, robaron, asaltantes, etc. Como se puede observar en la Figura 5.2 estas palabras son las más frecuentes en las noticias clasificadas en este evento delictivo.

Figura 5.2: Gráfica de palabras distintivas en la clase asalto.

Las palabras con mayor frecuencia en el evento delictivo de secuestro, como se puede ver en la Figura 5.3 son: secuestro, delito, víctimas, rescate, violencia, secuestrada, mujer, etc. Estas palabras son representativas a este evento delictivo.

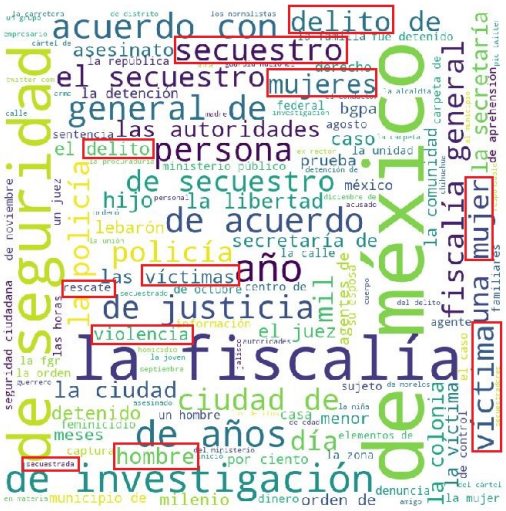

Figura 5.3: Gráfica de palabras distintivas en la clase secuestro.

Las palabras más representativas en violación se pueden observar en la Figura 5.4, estas palabras son: violación, víctima, abuso, abuso sexual, delito, violencia, niña, niño, etc. Todas estas palabras representan a este evento delictivo.

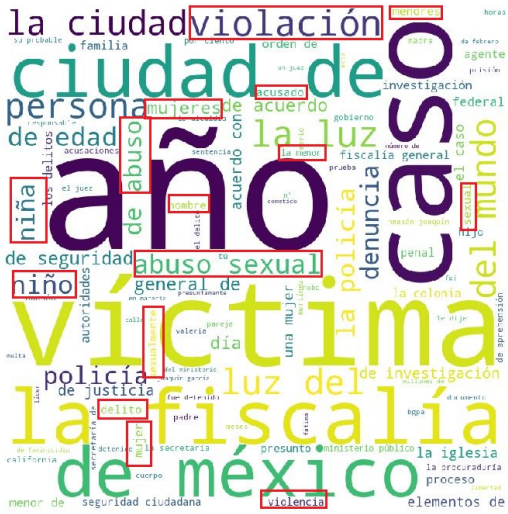

Figura 5.4: Gráfica de palabras distintivas en la clase violación.

En la Figura 5.5 podemos observar las palabras con mayor frecuencia en el evento delictivo de suicidio destacan las palabras como: suicidio, violencia, estudiante, pistola, víctima, vida, universidad, salud mental, depresión, etc. Estas palabras son representativas a este evento delictivo.

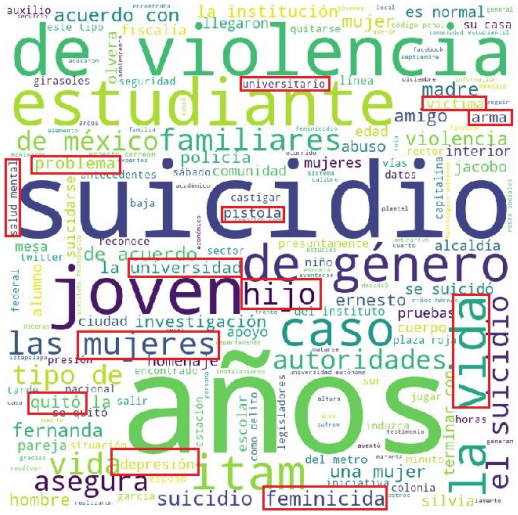

Figura 5.5: Gráfica de palabras distintivas en la clase suicidio.

Una vez analizadas las palabras más representativas en las noticias recabadas, se procede a realizar un análisis más profundo: este análisis consiste en la lectura de cada noticia, en donde se obtendrán las palabras más específicas que describen a cada una de las clases. Como se puede observar en la Figura 5.6 para la clase de homicidio las palabras más representativas son privar de la vida, matar, feminicidio, ultimar, asesinar, etc. Se pueden apreciar por clase cada una de estas palabras más representativas. Analizando estas palabras se observó que en la clase de suicidio se cuenta con un mayor vocabulario para poder identificar las palabras que indican que efectivamente es suicidio, además de que la mayoría de estas son oraciones, como, por ejemplo: se quitó la vida, termino con su vida, ya no quería seguir viviendo.

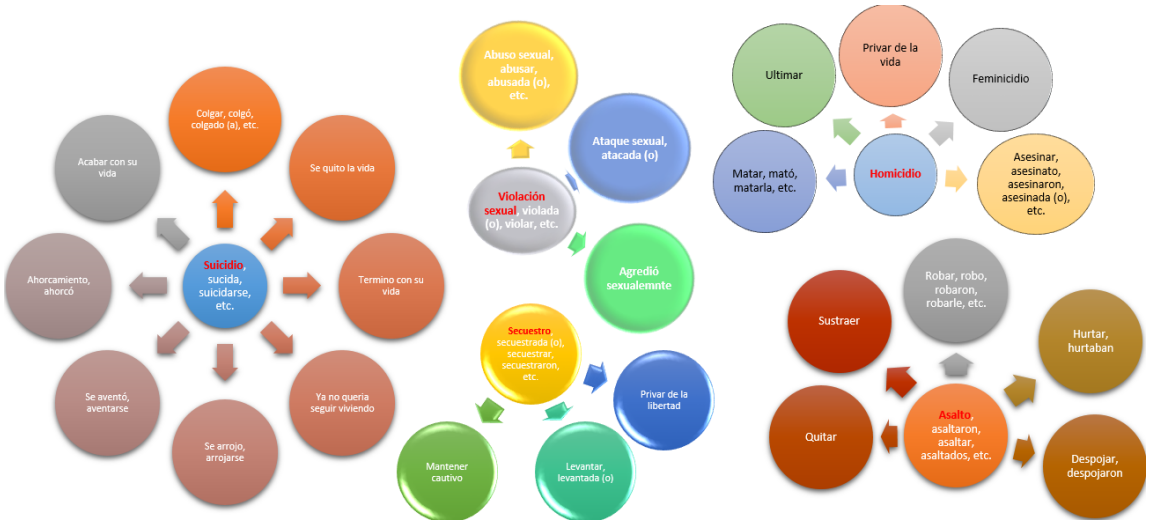

Figura 5.6: Gráfica de palabras distintivas de cada clase.

Posteriormente se procede a graficar los resultados obtenidos por estados, en donde se muestran los datos generales de los estados en donde suceden estos eventos delictivos en un periodo del 06 de agosto del 2019 al 23 de marzo del 2020, luego se muestran los estados por cada evento delictivo que vienen siendo: homicidio, secuestro, asalto, suicidio y violación. De ahí se muestran los datos de sexo por edades en cada evento delictivo, así como datos generales. Además de hacer un recuento de víctimas y culpables por sexo, estos se recuperan de las noticias clasificadas anteriormente ya que estas proporcionan estos datos. A continuación, se describen estas gráficas y se explica cada una.

En la Figura 5.7 se puede observar que la ciudad de México es la que cuenta con un mayor índice de eventos delictivos, siguiéndole Chihuahua y el Estado de México.

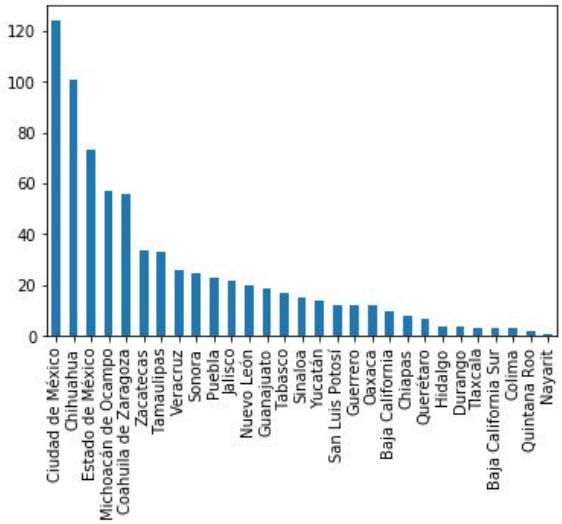

Figura 5.7: Gráfica de estados con mayor incidencia delictiva.

En los resultados obtenidos del análisis de las noticias previamente clasificadas, se procede a obtener el lugar en el que ocurre cada evento delictivo. En la Figura 5.8 se puede observar que Ciudad de México cuenta con el mayor número de homicidios, seguido del estado de Chihuahua, Estado de México, entre otros.

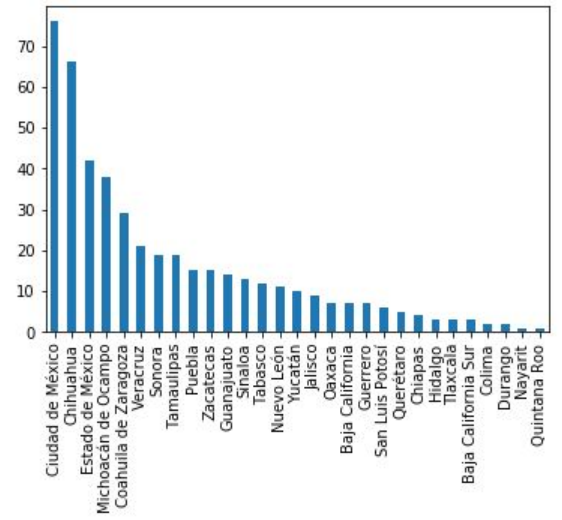

Figura 5.8: Gráfica de homicidio.

En la Figura 5.9 se puede observar que la Ciudad de México cuenta con el mayor número de asaltos, seguido de la Chihuahua, Estado de México, entre otros.

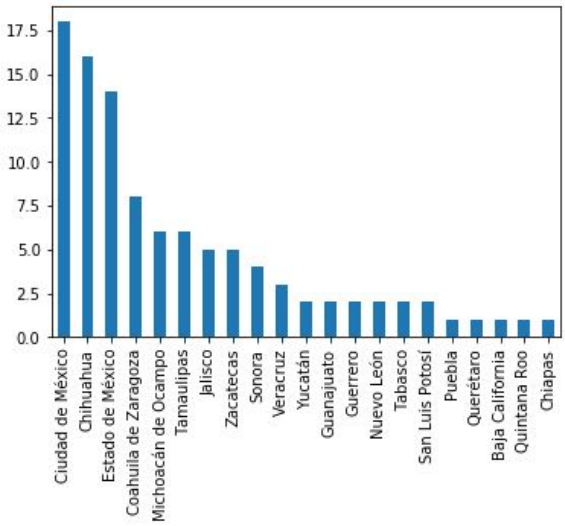

Figura 5.9: Gráfica de asalto.

En la Figura 5.10 se puede observar que la Ciudad de México cuenta con el mayor número de secuestros, seguido de Chihuahua, Coahuila, entre otros.

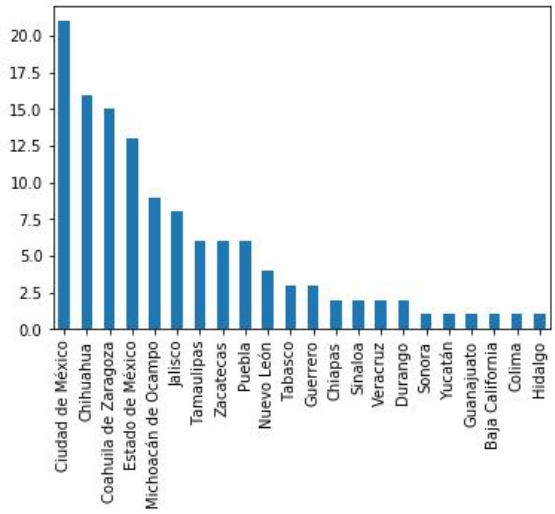

Figura 5.10: Gráfica de secuestro.

En la Figura 5.11 se puede observar que la Ciudad de México y Michoacán cuentan con el mayor número de suicidios.

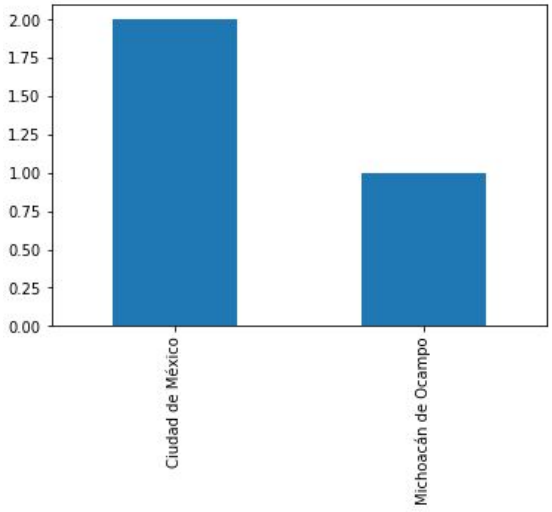

Figura 5.11: Gráfica de suicidio.

En la Figura 5.12 se puede observar que Zacatecas cuenta con el mayor número de violación sexual, seguido de la Ciudad de México, Oaxaca, entre otros.

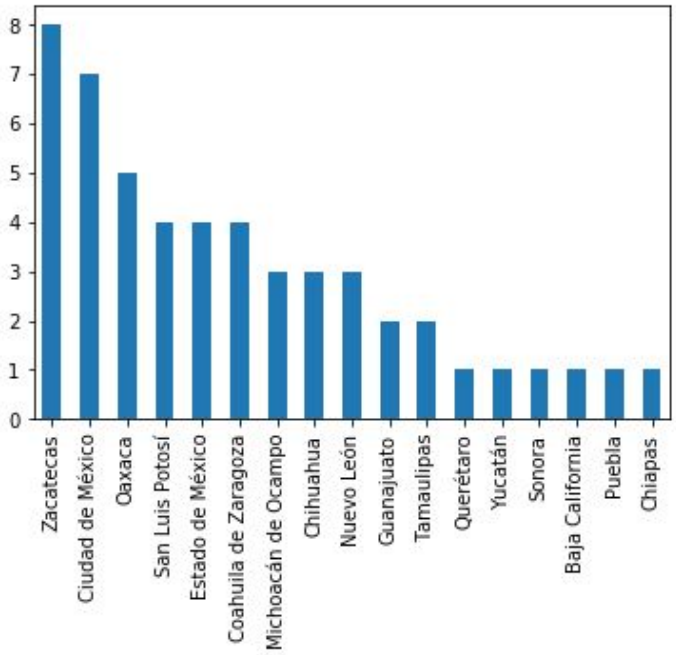

Figura 5.12: Gráfica de violación.

En los datos generales se observa que, en estos eventos delictivos las mujeres

cuentan con el mayor número de víctimas siendo un total de 520, como se puede observar en la Figura 5.13. También se presentan 329 víctimas, las cuales son hombres. Además se observa que 727 hombres y 73 mujeres son culpables en dichos eventos delictivos. A continuación, se presentan las edades correspondientes a víctimas y culpables por sexo.

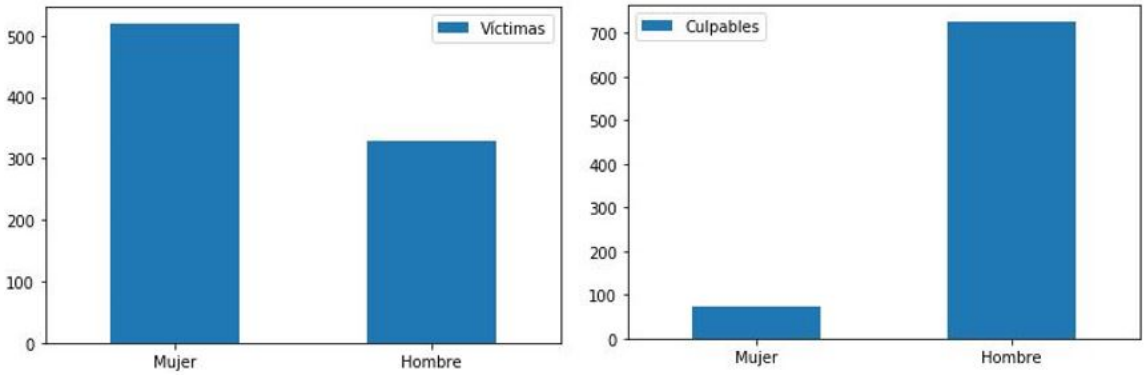

Figura 5.13: Gráfica de víctimas y culpables por sexo.

En la Figura 5.14 se aprecia que el rango de edades en las mujeres es de 1 a 84 años, siendo estas víctimas. Además de que hay más de 25 mujeres en una edad de 7 años las cuales son niñas, seguidas de jóvenes de 19 años con más de 10 al igual que 25 años, se presentan más de 5 casos en donde niñas de 1 año sufren de algún evento delictivo.

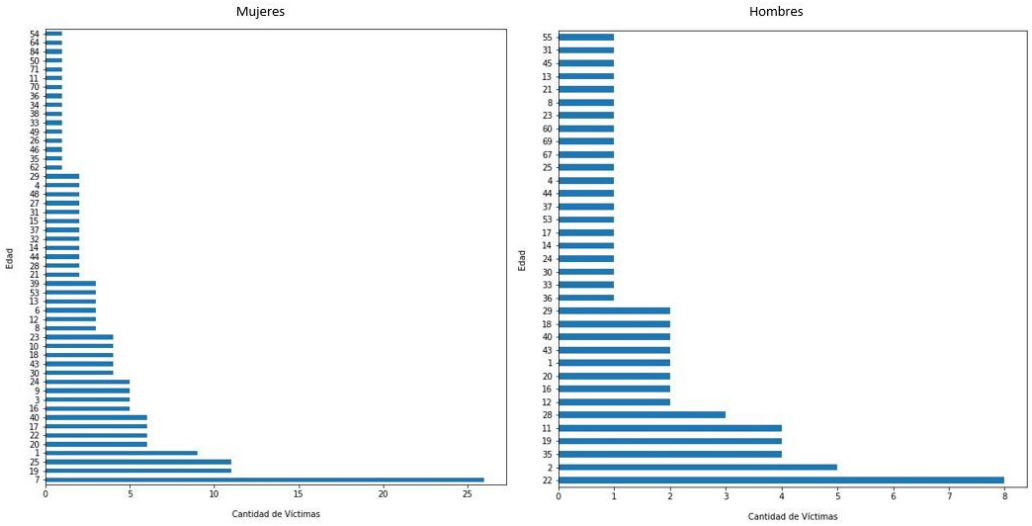

Figura 5.14: Gráfica de víctimas por sexo.

En la Figura 5.15 se muestran los resultados generales para mujeres y hombres culpables, siendo los hombres los que cuentan con un mayor número. En las edades de 16, 32 y 18 años hay un total de 8 culpables, seguidos de 6 hombres con edad de 19 años. Entre las mujeres son 41, 16 y 26 años con un total de 2 mujeres. Se observa claramente que los culpables son personas jóvenes los cuales cometen algún tipo de delito delictivo.

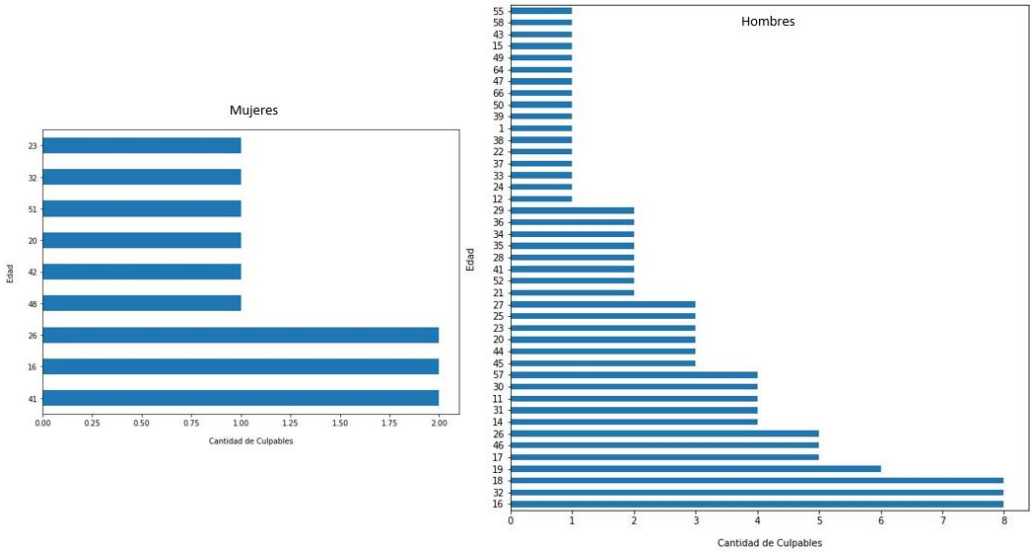

Figura 5.15: Gráfica de culpables por sexo.

En el evento delictivo de homicidio se puede observar en la Figura 5.16 en donde 357 mujeres y 247 hombres son víctimas. Además de que 36 son mujeres y 418 son hombres culpables de homicidio. Las mujeres siguen teniendo una taza mayor de victimización y los hombres en su mayoría son culpables de homicidio.

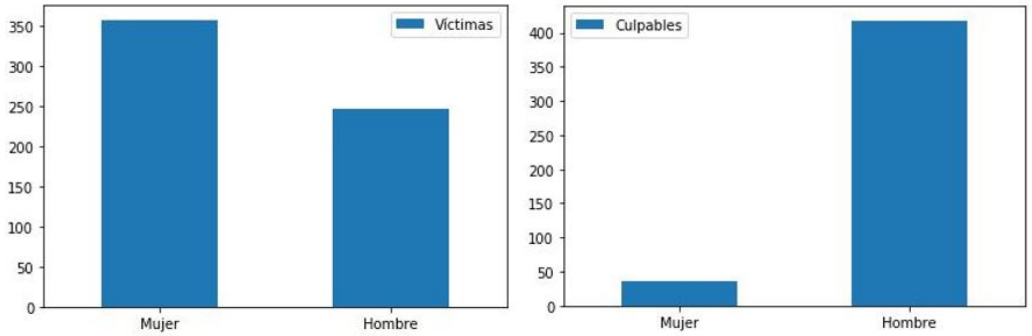

Figura 5.16: Gráfica de víctimas y culpables en homicidio.

Para las edades de hombre y mujeres víctimas en el evento delictivo de homicidio, se observa en la Figura 5.17 que más de 20 mujeres de edad de 7 años son víctimas, al igual que mujeres de edad de 25 años con más de 10 mujeres y con edad de 19 y edad de un año más de 8 mujeres sufren homicidio. Para los hombres hay 7 víctimas de 22 años, seguidas de 35, 2 y 11 años los cuales son asesinados.

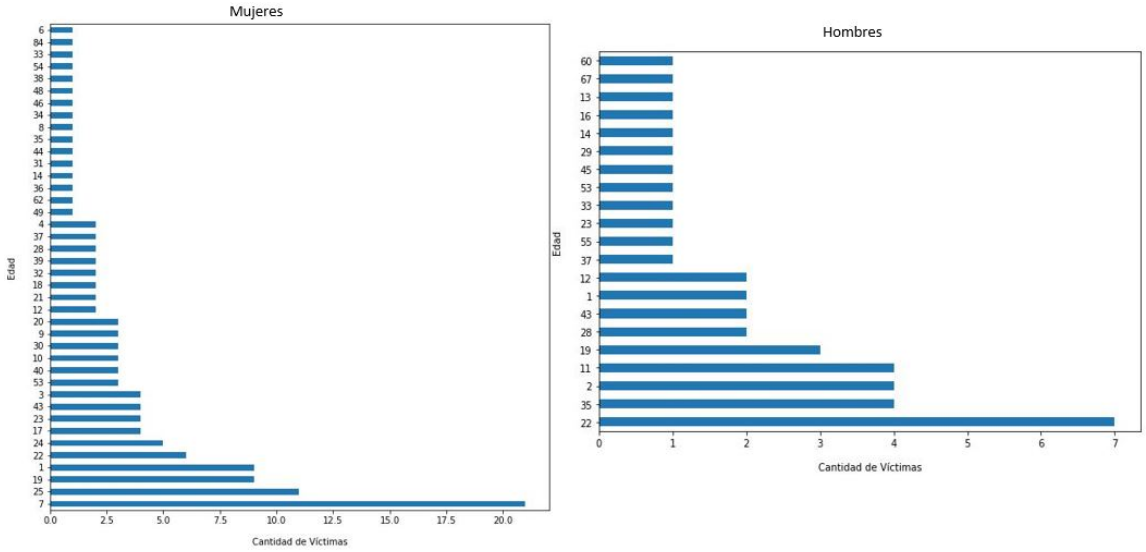

Figura 5.17: Gráfica de víctimas en homicidio.

Las estadísticas en el evento delictivo para los culpables en homicidio se presentan en la Figura 5.18 donde se presentan mujeres de 48 y 32 años, las cuales son culpables de homicidio. Con mayor número de culpables se encuentran los hombre, en la edad de 18 y 32 años hay 5 culpables, seguido de las edades de 46 y 11 años con 4 culpables para cada edad. Esto indica que hay menor número de mujeres culpables en homicidio y los hombres son los que más comenten este evento delictivo.

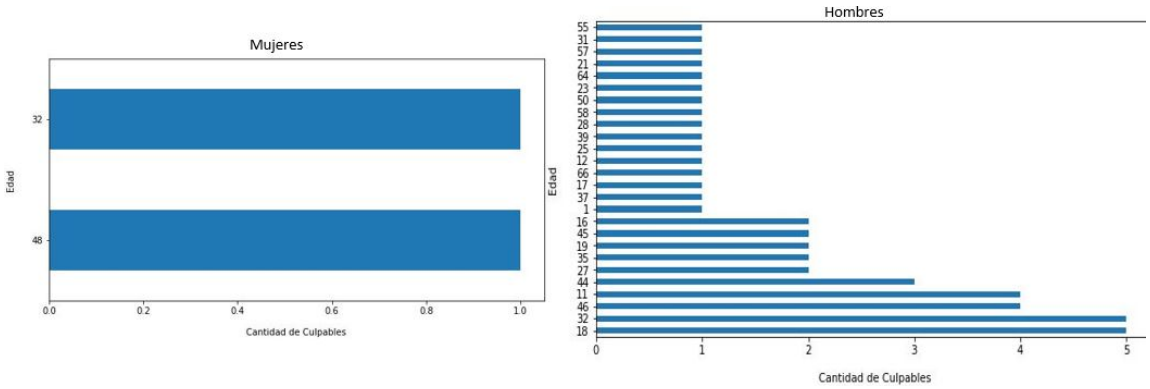

Figura 5.18: Gráfica de culpables en homicidio.

En el caso de asalto como se puede apreciar en la Figura 5.19 las víctimas son los hombres con un total de 34 y en menor cantidad las mujeres con 16 casos. Así mismo los hombres son los principales responsables de asalto con un total de 127 y con 18 mujeres siendo responsables de este evento delictivo.

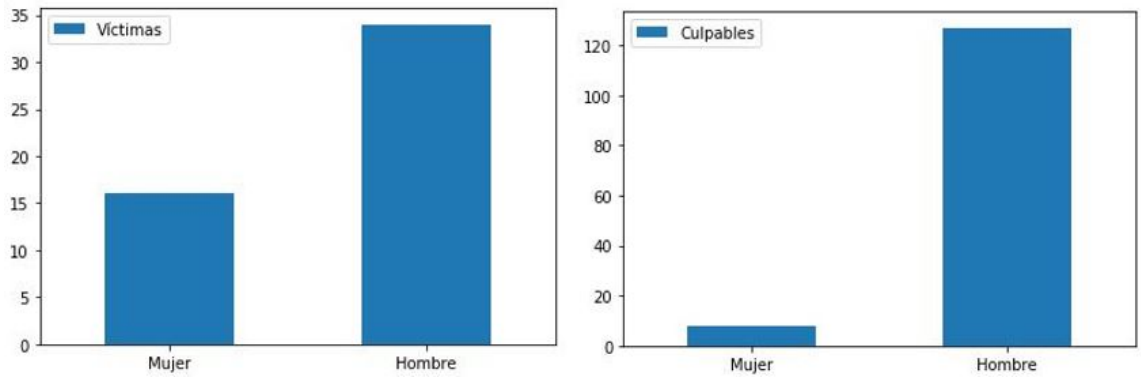

Figura 5.19: Gráfica de víctimas y culpables en asalto.

Entre las edades más comunes de víctimas de asalto de mujeres, es la edad de

29 años con dos víctimas, seguidas de 39, 40, 64 etc., en los casos de los hombres víctimas de asalto las edades ocurre en las edades de 18 y 40 años, seguidas de 69, 31, 24, etc. En este caso el evento delictivo de asalto sus víctimas son hombres jóvenes o adultos (ver Figura 5.20).

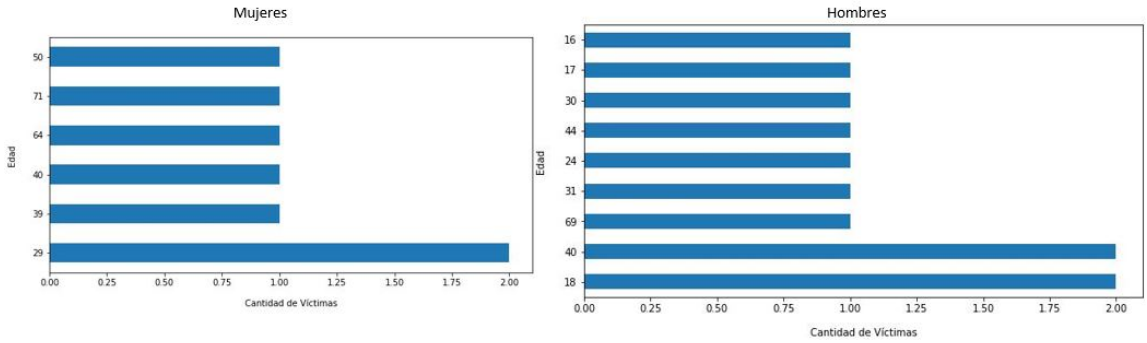

Figura 5.20: Gráfica de víctimas en asalto.

En la Figura 5.21 se muestran que los hombres son los principales responsables de asalto, entre las edades de 26 y 19 años hay 4 responsables de este acto delictivo, seguidos de 30, 18 y 32 años presentando a 3 responsables. En menor cantidad de culpables aparecen las mujeres, en una edad de 41 y 26 con 2 culpables.

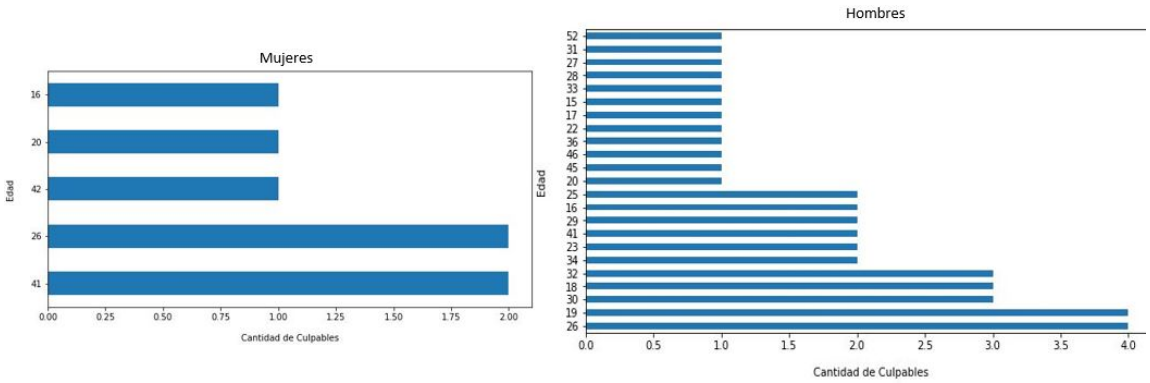

Figura 5.21: Gráfica de culpables en asalto.

Para el caso de secuestro en la Figura 5.22 siguen siendo víctimas, en mayor número las mujeres presentando 61 casos y 33 casos para los hombres, los cuales también han sido víctimas de este evento delictivo. Nuevamente los hombres son los

principales responsables de secuestro presentando 121 casos y con 21 culpables para mujeres.

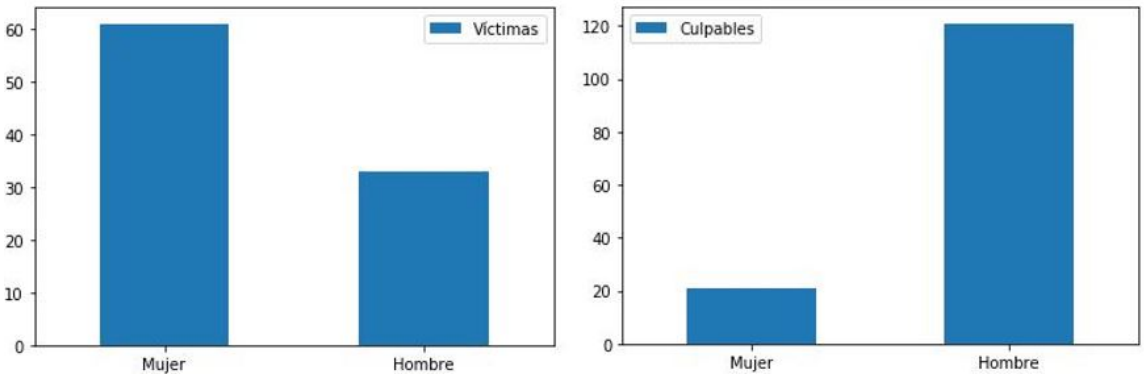

Figura 5.22: Gráfica de víctimas y culpables en secuestro.

Las edades de las mujeres que sufren de secuestro, son de 40 y 16 años presentándose dos casos y un caso para las edades de 18, 7, 20 años, etc. Se presentan además las víctimas, en este caso hombres de edades de 20, 36 y 29 años, con una ocurrencia por edad (ver Figura 5.23).

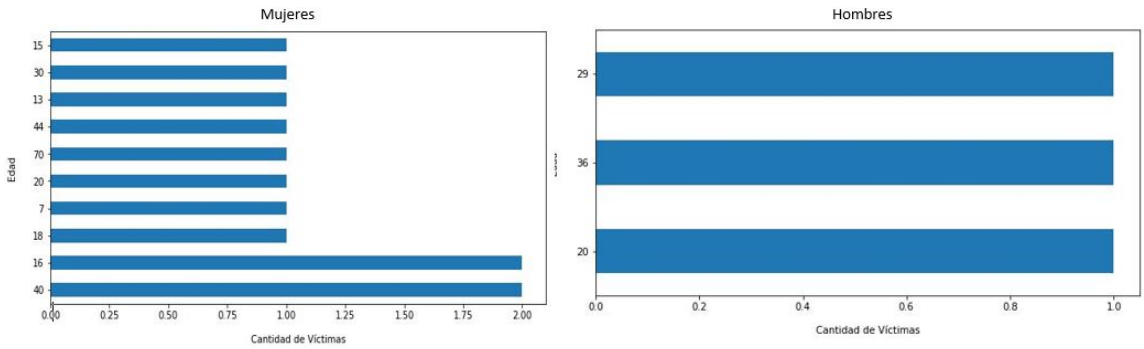

Figura 5.23: Gráfica de víctimas en secuestro.

En el caso de culpables, como se muestra en la Figura 5.24 hay 3 mujeres culpables en edades de 16, 51 y 23. Además se presentan 4 hombres culpables de edad de 16 y 14 años; con 3 caso se presentan en la edad de 17 años, los cuales son menores de edad.

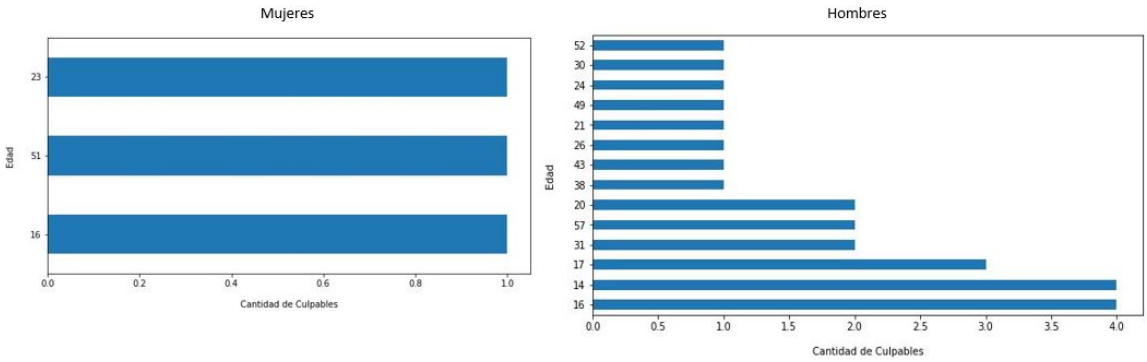

Figura 5.24: Gráfica de culpables en secuestro.

En el evento delictivo de violación se presentan en mayor número las mujeres quienes son más vulnerables a sufrir este tipo crimen, en este caso se presentan 74 víctimas, de igual manera los hombres sufren violación presentándose 5 casos. Para los culpables los hombres presentan 58 casos y las mujeres con 8 casos (ver Figura 5.25).

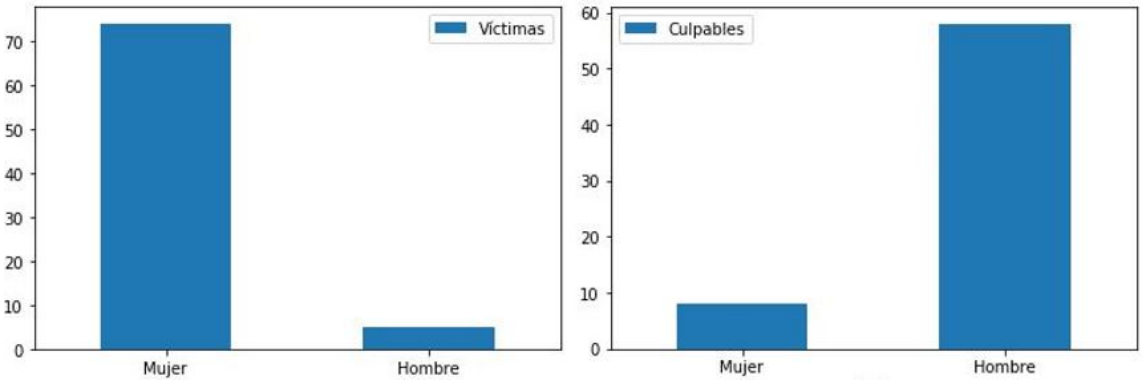

Figura 5.25: Gráfica de víctimas y culpables en violación.

En las edades en las que se presenta el mayor número de violación, en este caso son las mujeres de 7 años presentando 4 casos, seguidas de las edades de 16 años con 3 casos, como se muestra en la Figura 5.26. En los hombres se presenta un caso en edad de 20, 4 y 25 años.

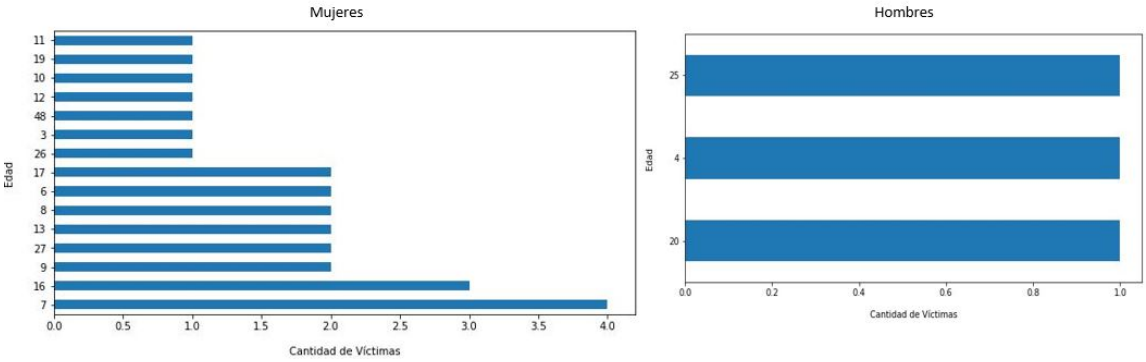

Figura 5.26: Gráfica de víctimas en violación.

En el caso de culpables de violación solo hay registro de hombres, los cuales cometen este delito, las edades de dichos hombres son de 36 y 57 años con un caso como se muestra en la Figura 5.27.

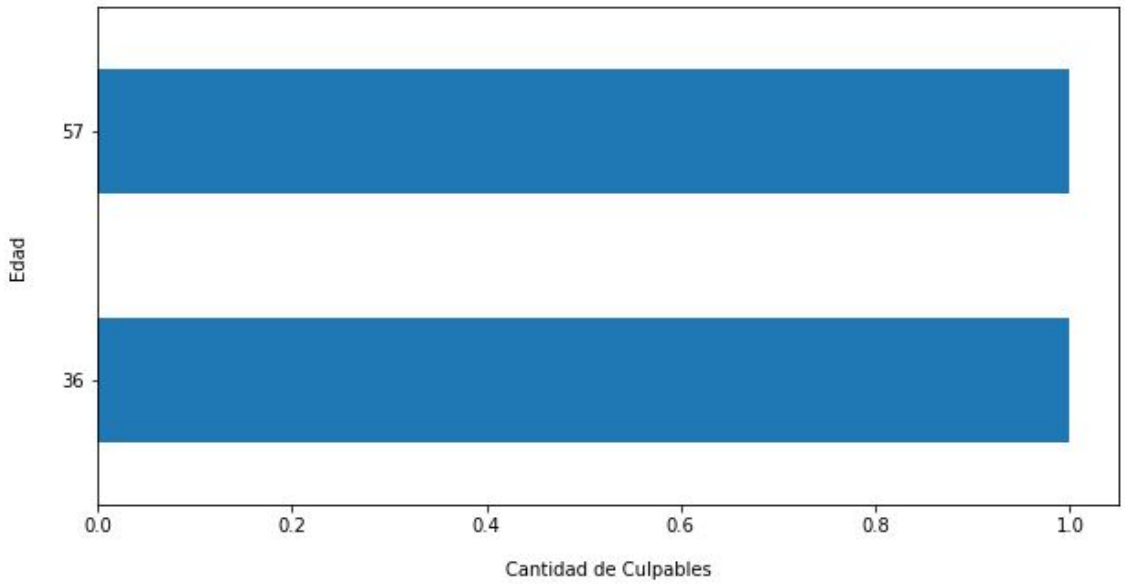

Figura 5.27: Gráfica de culpables en violación.

En el evento delictivo de suicidio se presentan con mayor incidencia en las mujeres con un total de 12 y 10 hombres víctimas, además se presentan 2 hombres culpables de suicidio, ya que son los que ocasionan este evento según los registros (ver Figura 5.28).

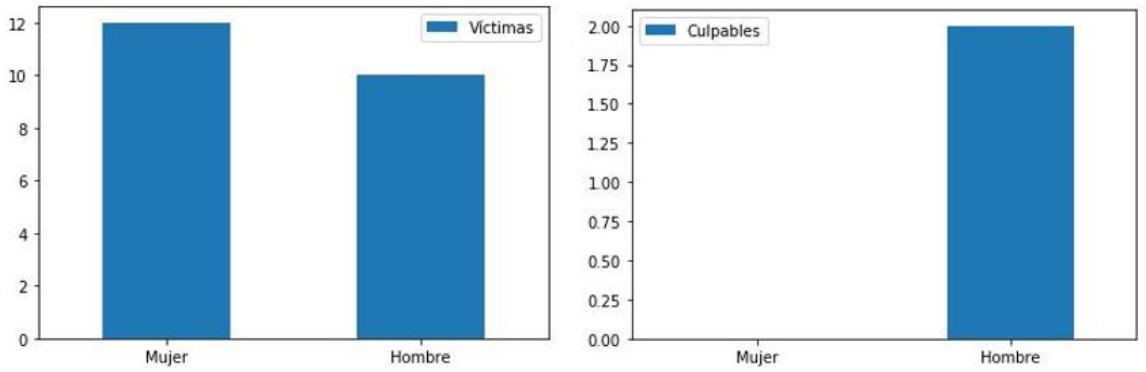

Figura 5.28: Gráfica de víctimas y culpables en suicidio.

Entre las edades registradas en el evento delictivo de suicidio en las mujeres se encuentras las de 20 años presentando 2 casos y con un caso las mujeres de edades de 14, 19, 18, 31 y 15 años en las que la mayoría son jóvenes las que sufren de suicidio. Para los hombres se presenta un caso en las edades de 22, 28, 21, 8, 19 y 2 años como se puede ver en la Figura 5.29.

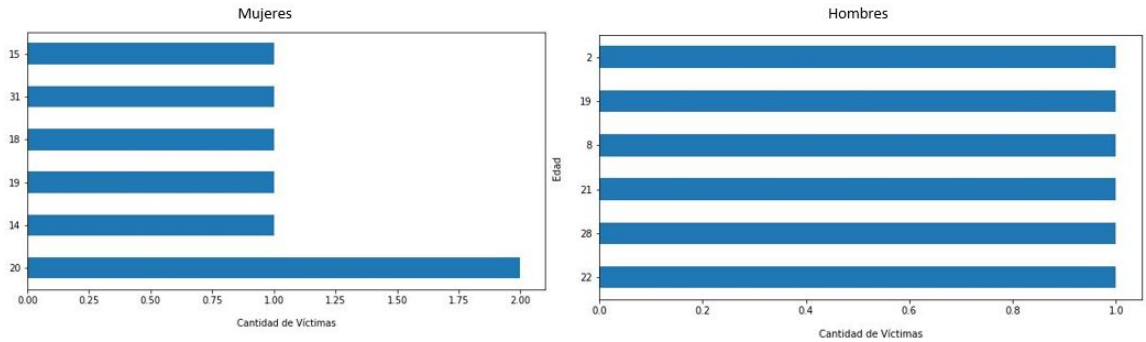

Figura 5.29: Gráfica de víctimas en suicidio.

Como se observó se obtuvo un buen resultado de clasificación, en la métrica de exactitud y la métrica F_1 arrojaron un 77 %. En la obtención de las palabras más representativas a cada evento criminal, los resultados son buenos, pues para cada evento se obtuvieron palabras referentes a homicidio, secuestro, asalto, suicidio y violación. En el caso de la detección de zonas de alto riesgo se observó que la Ciudad de México, Chihuahua y el Estado de México están entre los tres primeros lugares de mayor incidencia de actos delictivos en un lapso del 06 de agosto del 2019 al 23 de marzo del 2020. El acto delictivo que ocurre más en estos estados, principalmente es

el de homicidio y el que presenta menos incidencia es suicidio. Finalmente se pudo observar que en estos eventos delictivos las mujeres tienen el rol principalmente de víctimas y los hombres el rol de culpables, además de que las edades varían en cada evento delictivo.

Conclusiones

Para la realización de este trabajo de investigación se realizó un análisis de trabajos anteriores, los cuales se basaron en los temas de interés, para tener un punto de referencia y así iniciar el trabajo. Se desarrolló un algoritmo que consta de dos fases. En la primera fase se desarrolló un algoritmo para la clasificación automática de eventos delictivos en noticias periodísticas basado en la red CNN y la red LSTM. En la segunda fase se desarrolló un algoritmo para el análisis de las noticias. Para la primera fase se cuenta con dos corpus, el primer corpus se usa para realizar el entrenamiento, consta de encabezados de noticias obtenidas de la red social de Twitter obtenidas de [24] y para realizar las pruebas se cuenta con el segundo corpus anotado, las noticias de este corpus se obtuvieron de la página local: Milenio, en un periodo del 06 de agosto del 2019 al 23 de marzo del 2020. En base a los resultados experimentales en la fase de clasificación se obtuvo una exactitud global del 77 %, además de que la métrica F_1 global arrojó el mismo resultado, con la métrica global de precisión se obtuvo un 73 % y en la exhaustividad global se obtuvo un 62 %.

En la fase del análisis de las noticias clasificadas, se realizó un estudio por cada evento criminal para obtener las palabras más representativas de estos. Los resultados obtenidos son adecuados pues estas palabras son significativas para cada evento. Posterior a este análisis se recabaron las palabras u oraciones que describen cada evento criminal, por ejemplo, para la clase de homicidio se obtuvo una exactitud y una exhaustividad del 86 %, así como una precisión del 99 % y un F_1 del 92 %, siendo este el resultado más alto puesto que cuenta con un mayor número de noticias, además de que en estas se encontraron como palabras representativas: homicidio, feminicidio, asesinar, privar de la vida, etc. Por el contrario para la clase de suicidio se obtuvo una exactitud y una exhaustividad del 31 %, así como una precisión del 99 % y un F_1

del 48 %, en este caso es la clase con menor número de noticias, además de que las palabras que describen a este evento la mayoría son oraciones, por ejemplo: se quitó la vida, termino con su vida, acabo con su vida, ya no quería seguir viviendo, etc. Estos resultados nos dan una clara idea de que entre mayor sea el número de vocabulario en las noticias se obtendrán mejores resultados de clasificación por clase, además de que al momento de realizar la clasificación las redes neuronales, para realizar dicha clasificación se basan en las palabras para entrenar la red y poder proporcionar una adecuada clasificación, uno como persona al momento de leer las oraciones que se mencionan en la clase de suicidio, claramente sabremos que es suicidio, pero la maquina necesita de un mayor vocabulario en donde se mencionen estas palabras y así poder tener un buen entrenamiento y como resultado una buena clasificación.

Posteriormente se obtuvieron los estados con mayor frecuencia en cada evento, obteniendo que Ciudad de México, Chihuahua y Estado de México presentan el mayor número de estos eventos delictivos. Para el caso de de los eventos delictivos de homicidio, asalto y secuestro están entre los primeros dos lugares los estados de la Ciudad de México y Chihuahua. En el evento de suicidio esta la Ciudad de México y Michoacán de Ocampo y finalmente en el evento de violación sexual esta Zacatecas y Ciudad de México. El estado que está presente en todos estos eventos delictivos es la Ciudad de México, este es un dato muy importante para las autoridades y para la comunidad en general, pues se deben de tomar medidas más estrictas en este estado. En el análisis de víctimas y culpables se puede observar que las mujeres son las principales víctimas de los eventos de homicidio, secuestro, violación y suicidio, el caso de asalto las principales víctimas son hombres. En el caso de culpables los hombres cumplen ese rol en todos los eventos delictivos. Respecto a las edades, en la mayoría de estos eventos las víctimas suelen ser mujeres menores a 25 años y para el caso de mujeres culpables suelen ser mujeres jóvenes o adultas. En el caso de hombres víctimas están entre menores de 28 años en el evento de violación y suicidio, para el evento de homicidio, secuestro y violación suelen estar entre jóvenes y adultos, en el caso de culpables se encuentran jóvenes y adultos. Como se pudo observar en las edades hay variedad para las víctimas, así como para los culpables, lo que sigue siendo alarmante es que las mujeres siguen siendo un grupo vulnerable para estos eventos delictivos, así como los hombres culpables.

Como trabajo futuro se propone agregar un nuevo evento delictivo, el cual sería el de Narcóticos, pues en el etiquetado de noticias se encontró demasiado contenido sobre este tema, además de que sería un dato interesante para obtener las zonas de riesgo, además de un análisis de este evento. Otra propuesta es la clasificación de múltiples etiquetas, en donde se podría utilizar una red neuronal LSTM, además de otras. En esta clasificación se tomará la noticia y posteriormente se realizará la clasificación en varias clases a la que pertenezca. Finalmente se propone la búsqueda de patrones para la obtención automática de víctimas y culpables, así como sus edades. Además de buscar en la literatura alguna otra forma de abordar este problema.

Bibliografía

[1] scikit learn. https://scikit-learn.org/stable/modules/model_evaluation.html, 2007 - 2019. Accedido 20-12-2019.

[2] Municipios de méxico. http://www.municipios.mx/, 2020. Accedido 15-01-2020.

[3] wordcloud. https://pypi.org/project/wordcloud/, May 2, 2020. Accedido 03-02-2020.

[4] Paul David Cumba Armijos. *Predicción de ataques de cyber bullying mediante técnicas de aprendizaje profundo apoyandose en un corpus de entrenamiento para la clasificación de texto en español*. Tesis de maestría, Universidad Internacional SEK, 2018.

[5] beautifulsoup. `https://pypi.org/project/beautifulsoup4/`. Oct 6, 2019.

[6] Juan Marcelo Alvarado Carrera. *Clasificación de sílabos académicos en base a redes neuronales de aprendizaje profundo*. Grado en ingeniería, Universidad del Azuay, 2018.

[7] Juan Pablo Cárdenas, Gastón Olivares, y Rodrigo Alfaro. Clasificación automática de textos usando redes de palabras. *Revista Signos*, 47(86):46–364, 2013. ISSN 0718-0934.

[8] Li Deng y Dong Yu. *Deep Learning Methods and Applications*, tomo 7. Foundations and Trends in Signal Processing, 2013. ISBN 1932-8346. doi: 10.1561/2000000039.

[9] Carlos Escolano y Marta R. Costa-jussà. Generación morfológica con algoritmos de aprendizaje profundo integrada en un sistema de traducción automática

estadística. *Procesamiento del Lenguaje Natural*, 59:107–114, 2017. ISSN 1135-5948.

[10] Juan Diego Gómez Fierros. *Poblado automático de Ontologías Espaciales a Partir de Texto no Estructurado*. Maestría, Centro Nacional de Investigación y Desarrollo Tecnológico, Departamento de Ciencias Computacionales, 2012.

[11] Pau Agustín Granell. *Redes Neuronales Recurrentes: Una aplicación para los mercados bursátiles*. Grado en estadística, Universidad Politécnica de Catalunya, Junio 2017.

[12] John Hunter, Darren Dale, Eric Firing, Michael Droettboom, y el equipo de desarrollo de Matplotlib. Matplotlib. https://matplotlib.org/, 08 de abril de 2020. Accedido 20-01-2020.

[13] Lilian Chapa Koloffon, Leonel Fernández Novelo, y Sandra Ley. Prevención del delito en méxico: ¿dónde quedó la evidencia? méxico. México Evalúa, México, 2014. URL https://www.mexicoevalua.org/wp-content/uploads/2014/01/donde_quedo_la_evidencia.pdf.

[14] Verónica Lucía y Chamorro Alvarado. *Clasificación de tweets mediante modelos de aprendizaje supervisado*. Tesis de maestría, Universidad complutense de Madrid, 2018.

[15] García Gutiérrez Álvaro. *Machine Learning en Bases de Datos de Lenguaje Natural*. Tesis de maestría, UAM. Departamento de Ingeniería Informática, 2016.

[16] R. Montañés, R. Aznar, y R. Del Hoyo. Aplicación de un modelo hibrido de aprendizaje profundo para el análisis de sentimiento en twitter. *CEUR Workshop Proceedings 2172 (2018)*, pág. 3027–3036, 2018. ISSN 1613-0073.

[17] Antonio Moreno, Eva Armengol, Javier Béjar Lluís Belanche, Ulises Cortés, Ricard Gavaldà Juan Manuel Gimeno, Beatriz López, y Mario Martín Miquel Sànchez. *Aprendizaje automático*. Edicions UPC, Barcelona, 2019. ISBN 84-7653-460-4.

[18] David Nadeau y Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1), 2007.

[19] Ángel Javier Alonso Hernández. *Deep Learning aplicado al resumen de textos*. Doble grado, Universidad Complutense de Madrid, 2018.

[20] Oinkina y Hakyll. Understanding lstm networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/, August 27, 2015. Accedido 20-11-2019.

[21] Juan Sebastian Gelvez Prieto. Redes neuronales convolucionales y redes neuronales recurrentes en latranscripción automática. *research gate*, 2019.

[22] José Manuel Rodríguez Rama. *Aplicación de técnicas de machine learning a la detección de ataques*. Tesis de maestría, 2017.

[23] Ariel E. Repetur. *Redes Neuronales Artificiales*. Tesis de licenciatura, Universidad Nacional del Centro de la Provincia de Buenos Aires, 2019.

[24] José A. Reyes-Ortiz y Maricela Bravo. Enhancing patterns with linguistic information for criminal event recognition. *Journal of Intelligent and Fuzzy Systems*, 34:3027–3036, 2018.

[25] Javier Di Deco Sampedro. *Estudio y aplicación de técnicas de aprendizaje automático orientadas al ámbito médico: estimación y explicación de predicciones individuales*. Tesis de maestría, Universidad Autónoma de Madrid, 2012.

[26] Lilian Judith Sandoval. Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista tecnológica*, (11), 2018.

[27] scrapy. `https://scrapy.org/`.

[28] Daniel Alejandro Torres Silva. *Diseño y aplicación de una metodología para análisis de noticias policiales utilizando minería de textos*. Tesis de maestría, Universidad de Chile, 2013.

[29] spacy. `https://spacy.io/`. 2016-2019.

[30] SSP y INEGI. Clasificación estadística de delitos 2012. 2013.

[31] Rachael Tatman. Pre-trained word vectors for spanish. Kaggle, 2019. URL https://www.kaggle.com/rtatman/pretrained-word-vectors-for-spanish.

[32] Alberto Téllez Valero. *Extracción de Información con Algoritmos de Clasificación*. Tesis de maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica, México, 2005.

[33] Mari Vallez y Rafael Pedraza-Jimenez. El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines. 1. 2007. URL http://www.hipertext.net.

[34] Ervin Varga. *Practical Data Science with Python 3: Synthesizing Actionable Insights from Data*. 2019.

[35] Augusto Cortez Vásquez, Hugo Vega Huerta, y Jaime Pariona Quispe. Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2), 2009.

[36] Wikistat. Neural networks and introduction to deep learning. 2016. URL http://wikistat.fr/pdf/st-m-hdstat-rnn-deep-learning.pdf. [En línea: esta página está disponible el 21 de enero de 2016].

[37] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, y Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 2018. ISSN 1869-4101. doi:10.1007/s13244-018-0639-9. URL https://doi.org/10.1007/s13244-018-0639-9.

[38] Bárbara Yuste. Las nuevas formas de consumir información de los jóvenes. *Revista de estudios de juventud, junio 15*, (108), 2015.