INSTITUTO TECNOLÓGICO SUPERIOR DE TEZIUTLÁN

Tesis

"Sistema basado en técnicas de Procesamiento de Lenguaje Natural para la detección de efectos adversos en la salud por el uso de medicamentos"

PRESENTA:

MARIANO GIBRAN MONTERO COLIO

CON NÚMERO DE CONTROL

21TE0005P

PARA OBTENER EL GRADO ACADÉMICO DE:

MAESTRA EN SISTEMAS COMPUTACIONALES

CLAVE DEL PROGRAMA ACADÉMICO

MPSCO-0127

DIRECTOR (A) DE TESIS:

DRA. MARÍA DEL PILAR SALAS ZÁRATE

"La Juventud de hoy, Tecnología del Mañana"

TEZIUTLÁN, PUEBLA, JUNIO 2023

AGRADECIMIENTOS

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por la beca otorgada para realizar estudios de posgrado a nivel maestría a través de la convocatoria "BECAS NACIONALES PARA ESTUDIOS DE POSGRADO 2022".

DEDICATORIA PERSONAL

En el proceso de culminación de esta tesis, deseo expresar mi profundo agradecimiento a todas aquellas personas que contribuyeron de manera significativa para hacer posible este logro.

En primer lugar, quiero extender mi gratitud a mi directora de tesis la Dra. María del Pilar Salas Zárate, por su guía constante, apoyo incondicional y certeros consejos a lo largo de este viaje académico. Su experiencia y dedicación han sido fundamentales en la dirección de este trabajo, y estoy sinceramente agradecida por la oportunidad de aprender de ella.

Asimismo, quiero expresar mi reconocimiento a los profesores que me impartieron clases a lo largo de la maestría. Sus enseñanzas y conocimientos compartidos han contribuido en gran medida a mi formación y al desarrollo de este trabajo de investigación.

Finalmente, quiero agradecer a mi madre Luz del Carmen Colio Aburto y a mi padre Rafael Montero Colio; amigos Gustavo, Citlalli, Enrique; y a mi pareja Daniela por su constante aliento, comprensión y paciencia durante este proceso. Sus palabras de ánimo y amor han sido un motor fundamental para superar los desafíos que se presentaron en el camino.

En resumen, este logro no habría sido posible sin el apoyo y contribución de todas estas personas e instituciones. Cada uno de ustedes ha dejado una huella muy importante en mi camino académico y personal, y por ello, les expreso mi más sincero agradecimiento.

¡Gracias a todos!

Atentamente,

Mariano Gibran Montero Colio

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CARTAS DE ACEPTACIÓN

RESUMEN

La Organización Mundial de la Salud indica que los efectos adversos causados por medicamentos son de gran relevancia ya que cada persona tiene diversas reacciones sin importar si la dosis o tratamiento son los correctos y esto puede afectar a su salud. En el estado del arte se encuentran muy pocas metodologías para detectar de manera temprana estas reacciones adversas causadas por medicamentos (RAM) específicamente en el idioma español, además las metodologías existentes lo abordan desde registros médicos dejando de lado una gran fuente de información como lo son los comentarios en redes sociales. Es por lo que en este trabajo se presenta la creación de un corpus en español con datos de Twitter y Facebook y la realización de experimentos para la evaluación del corpus con un algoritmo de aprendizaje máquina SVM y dos modelos de redes neuronales RAM e IAN. Obteniendo los mejores resultados con un 0.86 de *Accuracy* con la red neuronal RAM. Así mismo se realizó la prueba del concepto mediante un caso de uso con el medicamento Metformina de la aplicación web cumpliendo así los objetivos planteados. Por último, se concluye que el desarrollo de un sistema para brindar información sobre los efectos adversos causados por medicamentos es de gran utilidad para las personas, además de aportar una contribución de calidad al estado del arte actual.

INTRODUCCIÓN

El presente documento se estructura de la siguiente manera; en el capítulo 1, se abordan los conceptos que soportan el desarrollo de esta investigación. Así mismo, se describe la problemática a resolver, que es la detección de efectos adversos mediante el uso de técnicas de PLN. Con el fin de aportar una herramienta de farmacovigilancia que brinde información a las personas sobre los efectos adversos reportados oficialmente como aquellos que han sido reportados por personas en comentarios de redes sociales. Así mismo se presentan los objetivos de esta investigación y los alcances y limitaciones a cubrir.

En el capítulo 2, se describen los trabajos relacionados al desarrollo de esta tesis. Este es un apartado vital para esta investigación ya que las investigaciones realizadas previamente y vigentes en el estado del arte aportan gran información sobre cuáles son los próximos pasos a seguir, que técnicas dan los mejores resultados y que áreas de oportunidad existen respecto a la identificación de efectos adversos. Con el análisis del estado del arte se identificó una carencia en el uso de datos de redes sociales en español para la construcción de corpus de efectos adversos causados por medicamentos. Así mismo las técnicas identificadas para la construcción de un modelo que identifique estos efectos adversos, son SVM y las redes neuronales RAM e IAN.

En el capítulo 3, se describe el desarrollo de la creación de un corpus con datos de redes sociales y el desarrollo del modelo predictivo utilizando la metodología CRISP-DM, siguiendo sus seis etapas de desarrollo. Igualmente, se describe el uso de la metodología SCRUM para el desarrollo del sistema que implementa la información recolectada con la metodología CRISP-DM, mediante una arquitectura dividida en capas con una capa de presentación que es una aplicación web y móvil, una capa de dominio con la creación de APIs REST y finalmente una capa de datos con una base de datos no SQL.

En el capítulo 4, se realiza el estudio y análisis de resultados, mediante la evaluación de las técnicas utilizadas para la creación del modelo predictivo. Así mismo se presenta el caso de estudio del medicamento Metformina, el cual es utilizado para el tratamiento de la diabetes, mediante el uso de la aplicación web desarrollada y las características de la misma. Además, se presenta la comprobación de la hipótesis.

Finalmente, en el capítulo 5, se presentan las conclusiones obtenidas de desarrollo de esta investigación, con un resultado positivo al cumplir con los objetivos planteados. También, se presentan recomendaciones para mejorar no solo el modelo sino el sistema desarrollado y algunos trabajos a futuro que pueden desprenderse del desarrollo de esta investigación.

CAPITULO I GENERALIDADES DEL PROYECTO

1.1 Marco teórico

1.1.1 Procesamiento de Lenguaje Natural

Por Procesamiento de Lenguaje Natural (PLN, denominado también NLP por sus siglas en inglés) se entiende como la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje natural (Gelbukh, 2010). PLN utiliza y formula mecanismos para establecer la comunicación entre personas y máquinas de manera computacional, comúnmente el lenguaje natural se analiza en distintos niveles los cuales son:

- Análisis morfológico, se extraen las raíces, rasgos reflexivos y unidades léxicas compuestas de las palabras.
- Análisis sintáctico, se analiza la estructura sintáctica de una frase de acuerdo con la gramática aplicable.
- Análisis semántico, extracción del significado de una frase.
- Análisis pragmático, análisis profundo del significado más allá de la frase.

Entre los mecanismos aplicables al análisis del lenguaje mediante PLN, se utilizan principalmente técnicas lingüísticas formales basas en reglas de estructura aplicadas a las fases del análisis y técnicas probabilísticas que se basan en conjuntos de textos de referencia con características probabilísticas asociadas a las distintas fases de análisis (Martín Mateos & Ruiz Reina, 2013).

1.1.2 Análisis de Sentimientos/Minería de Opiniones

El análisis de sentimientos, también llamado minería de opiniones, es el campo de estudio que analiza las opiniones, los sentimientos, las evaluaciones, las valoraciones, las actitudes y las emociones de las personas hacia entidades como productos, servicios, organizaciones, individuos, asuntos, eventos, temas y sus atributos (Liu, 2012). De esta manera, aplicando diversas técnicas para el análisis de sentimientos se puede determinar si una opinión, texto o frase tiene una connotación positiva o negativa de acuerdo con el contexto.

El análisis de sentimientos brinda múltiples beneficios entre los que se encuentran, el poder obtener la valoración de opiniones referentes a productos o servicios, corrección de opiniones, mejora de sistemas de recomendación, conocer la valoración política, poder analizar el mercado financiero, entre muchas otros. Así mismo, el análisis de sentimientos ha aumentado en la última década debido al incremento del uso de métodos de aprendizaje automático y su uso dentro del PLN,

la gran disponibilidad de datos en internet gracias a la Web 2.0 y la creciente necesidad de las empresas por aprovechar la información que generan sus clientes (Pauli, 2019).

En el análisis de sentimientos se pueden distinguir tres niveles de profundidad los cuales son:

- Análisis a nivel de documento, se analiza de manera global un documento como un todo, clasificándolo como positivo, negativo o neutro, según se haya definido en el sistema de clasificación. Es un nivel superficial ya que la valoración solo aplica a una única entidad contenida dentro del documento.
- Análisis a nivel de oración, en este nivel se analizan las oraciones de un documento o texto y a cada una se le puede aplicar una valoración específica.
- Análisis a nivel de aspecto y entidad, es el análisis con mayor profundidad, en donde una entidad está formada por distintos elementos y sobre estos elementos se puede emitir una opinión y valoración propia.

1.1.3 Web Scraping

Web Scraping, también conocido como extracción o recolección de la Web, es una técnica para extraer datos de la Word Wide Web (WWW) y guardarlos en un sistema o base de datos para su posterior recuperación o análisis. Esto se realiza por un usuario o automáticamente por un bot o rastreador de la web. Debido al hecho de que la WWW constantemente genera una enorme cantidad de datos, esta técnica es eficazmente reconocida para la recolección de grandes volúmenes de datos (Zhao, 2017).

El Web Scraping es útil debido a que la mayoría de los datos que aparecen en las páginas web solamente pueden visualizarse a través de navegadores y no existe la posibilidad de guardar, copiar o modificar los datos para su uso de manera sencilla, perdiendo así datos valiosos para uso personal, investigativo o privado. Antiguamente, la única manera de extraer estos datos era de manera manual, lo cual implicaba mucho tiempo y esfuerzo por parte de una o más personas para recolectar una gran cantidad de datos. Usualmente un Web Scraper, se compone de dos partes un componente para la navegación y otro para la extracción de datos.

1.1.3.1 Web Crawling

Un Web Crawler, es un programa que se encarga de inspeccionar páginas web de manera automatizada y metódica utilizando algoritmos recursivos e inspeccionando los diversos enlaces y vínculos dentro de una página web, es por ello por lo que también se le conoce como *spider* o "araña web" ya se navega a través de la red

interna de una página web. Durante este proceso de inspección el web crawler descarga las páginas asociadas a las URL (*Uniform Resource Locator*, Localizador Uniforme de Recursos) dadas y finaliza su ejecución una vez ha completado su recorrido y descargado todas las páginas de la web inspeccionada o bien ha llegado a su límite de frontera definido.

Los Web Crawler sirven para distintos fines, entre los cuales se encuentra la aplicación en motores de búsqueda, indexadores de páginas web, creación de índices de consultas, encontrar páginas web de acuerdo con un contenido especifico, sistemas que reúnen corpus de páginas web, así como su aplicación para la minería de datos. Dado a la gran cantidad de datos existentes en las páginas web la aplicación de Web Crawler para reunir grandes cantidades de datos de manera eficaz y sencilla se ha tornado cada vez más importante tanto para investigaciones como para la toma de decisiones en la industria (Olston & Najork, 2010).

1.1.3.2 Web Information Extraction

Este es un programa que se encarga de extraer la información deseada de una página web tomando en cuenta que estas páginas tienen una estructura basada en HTML y que la información puede estar contenida en *tags* o "etiquetas". Así mismo, se debe identificar un patrón dentro de las páginas con el fin de utilizar selectores tales como *XPath o CSS Selectors* para finalmente almacenar la información en un formato definido por el usuario y acorde a sus necesidades (Stenhouse, 2017).

1.1.4 Corpus

Un corpus se puede definir como una colección de piezas del lenguaje que son seleccionadas y ordenadas de acuerdo con un criterio lingüístico explicito con el fin de ser utilizado como un ejemplo del lenguaje (Torruella & Llisterri, 1999). Así mismo el uso de corpus ha demostrado ser eficaz para encontrar la solución a algunos problemas tradicionales de la lingüística computacional, la traducción automática, entre otros.

Es importante hacer notar que existen distintos tipos de corpus, esto depende del grado de especificación de los criterios de selección para la recopilación de textos o de piezas del lenguaje. Entre las distinciones de corpus se encuentran las siguientes:

- Archivo o colección, conjunto de textos en soporte informático, sin relación aparente entre los mismos.
- Biblioteca de Textos Electrónicos, colección de textos en soporte informático, guardados en formato estándar siguiendo normas de contenido, pero sin un criterio riguroso de selección.

- Corpus Informatizado, recopilación de textos seleccionados de acuerdo con un criterio lingüístico, almacenados de manera estándar y de forma homogénea, destinado al análisis mediante procesos informáticos.

1.1.5 Base de Datos

Una Base de Datos (BD) se define como una representación de la realidad plasmada como un modelo que abarca solo una parte de la realidad dependiendo del contexto (Camps Paré et al., 2005). Así mismo para poder administrar una BD, es necesario contar con un Sistema Gestor de Bases de Datos (SGBD), cada SGBD puede ser utilizado para modelar BD dependiendo el tipo de BD. Por otro lado, todo modelo de BD proporciona tres tipos de herramientas básicas:

- Estructura de datos, define la construcción de tablas, objetos, componentes, árboles o entre otros.
- Reglas de integridad, son las restricciones que debe cumplir el SGBD a los datos.
- Operaciones, son la serie de procesos aplicables a los datos, tales como la recuperación de datos dada una sentencia, operaciones de actualización y eliminación de datos.

1.1.5.1 Relacionales

Modelo creado en 1970 por Codd en el cual se utilizan tablas para representar de manera lógica de los datos y la relación existente entre los mismos. Los elementos de los cuales se componen las tablas son las tuplas, las cuales representan a cada fila de la tabla y de campos o atributos que representan a cada columna de la tabla. Este modelo es de los más utilizados presente en Sistemas Gestores de Bases de Datos (SGBD) como Oracle, MySQL, SQL Server, PostgreSQL, entre otros (Piñeiro Gómez, 2014).

1.1.5.2 No Relacionales

Modelo creado para superar las limitaciones de las bases de datos relacionales y a su vez ofrecer soluciones más adaptadas a la computación distribuida o en la nube. Este modelo tiene características como el manejar una mayor cantidad de datos mayor a los modelos relacionales, no es necesario contar con una estructura previa para almacenar datos lo cual brinda una mayor flexibilidad, gran escalabilidad y diseño sencillo y de bajo coste (Alonso-Zárate & Casas Roma, 2021). Usualmente la manera de almacenar los datos se clasifica en cuatro tipos:

- Bases de datos clave y multivalor.
- Bases de datos orientadas a columnas.

- Bases de datos documentales.
- Bases de datos basadas en grafos.

1.1.6 Efectos Adversos de los Medicamentos

Un efecto adverso causado por medicamentos se puede definir como una reacción apreciablemente perjudicial o desagradable resultante de una intervención relacionada con el uso de un medicamento. Las reacciones adversas de un medicamento (RAM) se derivan principalmente de los datos de los ensayos clínicos y se aumentan mediante la vigilancia posterior a la comercialización (Lavertu et al., 2021a).

Entre las RAM se pueden distinguir dos tipos de reacciones una denominada reacción Tipo A, las cuales están relacionadas con las propiedades del medicamento producto de una reacción farmacológica de acuerdo con una dosis terapéutica habitual; y Tipo B, que son reacciones no farmacológicas del medicamento y por lo tanto impredecibles. Es por ello por lo que para diagnosticar y prevenir las RAM se deben tomar en cuenta factores de riesgo tales como antecedentes del paciente, exposición recurrente a fármacos o a otros medicamentos, situaciones clínicas específicas, la edad, entre otros factores (Porto Arceo, 2019).

1.1.6.1 Interacción Medicamento-Medicamento

La Interacción Medicamento-Medicamento (DDI) se define como la coadministración de dos o más fármacos al mismo tiempo y que puede afectar a la acción biológica de los fármacos implicados. La interacción puede afectar gravemente a los perfiles de eficacia y seguridad de los fármacos. Los principales tipos de DDI incluyen interacciones farmacocinéticas y farmacodinámicas (Vilar et al., 2018a).

1.1.7 Farmacovigilancia

La farmacovigilancia es un amplio espectro de actividades que se centran en identificar y prevenir los efectos adversos de los medicamentos, así como en comprender los factores de riesgo y las causas de las RAM cuando se producen (Chapman et al., 2019a). En los últimos años los alcances de la farmacovigilancia han ido aumento y hoy en día incluyen los siguientes:

- RAM o eventos adversos
- Errores de medicación
- Medicamentos falsificados o de calidad inferior
- Falta de efectividad de los medicamentos
- Uso indebido y/o abuso de medicamentos

- Interacción entre medicamentos

Existen sistemas de farmacovigilancia a nivel mundial dado a que es un tema de salud pública y es de vital importancia el poder notificar a las personas de los efectos adversos de un medicamento, esto se logra mediante la recopilación sistemática, el cotejo y el análisis de notificaciones o reportes de sospechas de RAM con el fin de gestionar los riesgos (Organización Mundial de la Salud, 2019).

1.2 Planteamiento del Problema

De acuerdo con la Organización Mundial de la Salud (OMS) (1946), la salud se define como un estado completo de bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades. La salud siempre ha sido uno de los temas más relevantes a nivel mundial, y esto se incrementó debido a la reciente pandemia causada por el virus SARS-COV-2, en la cual de acuerdo con estadísticas actuales han muerto más de 6 millones de personas y cerca de 519 millones se vieron afectadas de manera moderada a grave (Hale et al., 2021). Así mismo, existen múltiples factores que atentan en contra de la salud de las personas, uno de estos factores son las reacciones adversas causadas por medicamentos (RAM), que se puede definir como una reacción sensiblemente perjudicial o desagradable resultante de una intervención relacionada con el uso de un medicamento (Lavertu et al., 2021b) y es que alrededor del mundo países como Estados Unidos, Canadá y Japón gastan anualmente un aproximado de 500 mil millones, 65 mil millones y 7 mil millones de dólares respectivamente, todo esto para cubrir los gastos causados por las RAM.

Existen múltiples esfuerzos para mejorar la seguridad de las personas y mitigar los riesgos causados por las RAM, de esto se encarga la farmacovigilancia, a través de múltiples actividades, utilizando sistemas electrónicos de registro de salud de los pacientes, así como sistemas más avanzados que utilizan algoritmos de aprendizaje máquina basados en aprendizaje profundo o en redes neuronales convolucionales (CNN), los cuales suelen requerir de una infraestructura computacional lo suficientemente poderosa para ser entrenados (Chapman et al., 2019b). Debido a lo anterior el uso del procesamiento del lenguaje natural (PLN) se ha aprovechado como una herramienta para desarrollar sistemas rápidos, fácilmente entrenables y que no requieren de un hardware especializado (Chapman et al., 2019b), estos sistemas se alimentan de datos de las RAM obtenidos de textos de redes sociales (Twitter, Reddit, Facebook), foros relacionados con la salud y registros electrónicos de salud (I. S. Alimova & Tutubalina, 2020).

De acuerdo con lo descrito en los párrafos anteriores se pretende realizar un sistema para evaluar y notificar a las personas las reacciones adversas causadas por medicamentos con datos obtenidos de medios sociales, bases de datos de medicamentos y páginas de organizaciones oficiales, tales como, la de la Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS), utilizando técnicas de NLP para el idioma español. De esta manera será posible disminuir los gastos causados por las RAM y principalmente tener un impacto positivo en la salud de las personas que utilicen el sistema, al estar informadas sobre los riesgos de tomar un medicamento.

1.3 Justificación

La creación de herramientas que ayuden a ser más eficientes los sistemas de farmacovigilancia con el fin de tener presente las reacciones adversas causadas por medicamentes especialmente cuando se tratan de enfermedades crónicas tales como: diabetes mellitus tipo 2, hipertensión, entre otras, son de vital importancia, ya que brindan información confiable para las personas que padecen estas enfermedades. Es por lo que en este trabajo se propone el desarrollo de un sistema que implemente técnicas de minería de opiniones, PNL, minería de datos y Web Scraping con el fin de obtener comentarios de redes sociales, tales como Twitter y Facebook, en el idioma español en donde se mencionen efectos adversos causados por medicamentos para tratar estas enfermedades, realizar la identificación de las menciones y poder notificar a los usuarios sobre posibles efectos no reportados comparándolos contra información oficial, así como la obtención de documentos y alertas sanitarias por parte de la COFEPRIS.

Todo lo anterior da como resultado una herramienta Web y una aplicación móvil que permite a los pacientes conocer de manera fácil y rápida los posibles efectos adversos de un medicamento tanto los reportados oficialmente (RAM Tipo A) como los identificados por medio comentarios reportados por otros pacientes (RAM Tipo B), sus características y contraindicaciones, además de poder consultar la nueva información y puedan visualizar el resumen de las alertas sanitarias emitidas por la COFEPRIS, permitiendo a los pacientes tener una mayor información sobre los medicamentos que toman y evitar así complicaciones que pueden afectar de manera negativa a su salud y calidad de vida.

1.4 Hipótesis

La implementación de un sistema de software basado en técnicas de Procesamiento de Lenguaje Natural permitirá detectar en redes sociales efectos adversos en la salud causados por el uso de medicamentos.

1.5 Objetivo general

Desarrollar un sistema de software para la detección de efectos adversos en la salud por el uso de medicamentos a través de técnicas de Procesamiento de Lenguaje Natural y Big Data.

1.6 Objetivos específicos

- Realizar un análisis del estado del arte para identificar herramientas de software más utilizadas y las principales técnicas de Procesamiento de Lenguaje Natural y Big Data en el dominio de la salud.
- Identificar fuentes de información estructurada y no estructurada para la obtención de información de medicamentos
- Diseñar una arquitectura del sistema de farmacovigilancia.
- Desarrollar un módulo de extracción de información a través de fuentes de datos estructuradas y no estructuradas.
- Desarrollar un módulo de análisis de información a través de técnicas de PLN.
- Desarrollar e implementar el sistema que integre los módulos desarrollados.
- Realizar al menos un caso de estudio como prueba de concepto.

1.7 Alcances y limitaciones

Alcances

- El sistema software se desarrollará para el idioma español.
- El sistema software incluye un módulo de extracción de comentarios de medicamentos de la red social Twitter
- El sistema software incluye un módulo que realizará una monitorización constante de alertas sanitarias de medicamentos emitidas en México por la COFEPRIS a través de su sitio Web oficial.
- Como caso de estudio el sistema está orientado en medicamentos de enfermedades crónico-degenerativas, específicamente, diabetes e hipertensión, principales causas de muerte en México.

Limitaciones

- Carencia de herramientas de PLN para el idioma español.
- Carencia de banco de medicamentos en idioma español.
- Los comentarios expresados en medios sociales son difíciles de procesar debido a que contienen errores de ortografía y son más difícil de analizar a través de técnicas de PLN debido a la jerga utilizada.

CAPITULO II ESTADO DEL ARTE

Para comenzar con este capítulo se expone una breve síntesis de los trabajos que tienen una mayor relevancia sobre el tema de esta tesis y como estos aportan a la solución propuesta, así mismo se muestra el análisis realizado sobre los trabajos relacionados comparando los aspectos más relevantes de los mismos, debido a que esto da paso a las alternativas de solución y finalmente a la descripción solución propuesta con base al estado del arte actual.

2.1 Trabajos relacionados

A manera de introducción, es necesario conocer el panorama general de los diversos trabajos existentes acerca del SA (*sentiment analysis*, análisis de sentimientos) tal como lo presenta Zunic et al., (2020) realizaron un análisis sistemático sobre los trabajos relacionados a SA específicamente en la salud y bienestar con el fin de establecer de manera sistemática el conocimiento actual. En su estudio identificaron 86 estudios relevantes siguiendo la siguiente metodología: 1) Pregunta de investigación para definir el alcance; 2) Definición de un proceso de búsqueda; 3) Inclusión y exclusión de criterios para definir el alcance; 4) Valoración critica de los resultados de las revisiones; 5) Extracción de información relevante de los estudios, 6) Síntesis de los datos y síntesis de las pruebas para respaldar conclusiones de la revisión. Como resultados obtuvieron que MedHelp es el sitio web más utilizado para la extracción de datos debido que permite a sus usuarios compartir información acerca de experiencias personales y generación de información basada en evidencia con 298 temas acerca de la salud y el bienestar. Así mismo dentro de los artículos revisados encontraron cinco roles principales asociados a las opiniones de las personas, estos roles son 1) Enfermo, 2) Adicto, 3) Paciente, 4) Cuidador y 5) Victima de Suicidio, siendo que en 42 de 86 artículos se centran en datos e información generada por enfermos y pacientes. También establecen que los algoritmos de aprendizaje máquina más utilizados para SA relacionados al análisis de la salud y el bienestar son: 1) SVM (*suport vector machine*, máquina de soporte de vectores), 2) NB (*Naïve Bayes classifier*, Clasificador Naïve-Bayes) y 3) Aprendizaje de árbol de decisiones (*decision tree learning*).

Las RAM (*Adverse Drug Reactions,* Reacciones Adversas a Medicamentos) son un problema que ha ido creciendo a lo largo de los años en distintos países causando problemas graves de salud e inclusive la muerte en casos severos, es por lo que Liu et al. (2018) proponen un método para la identificación de RAM en foros de salud en línea en Estados Unidos. Este método se enfoca en minería de datos de pacientes y las experiencias de estos, primeramente, inician con la obtención de los datos del

foro de enfermedades cardiacas MedHelp en la cual recolectaron 120,275 entradas con un total de 998,637 oraciones, agrupadas en 34,065 hilos; posteriormente realizan un preprocesamiento mapeando los términos usados en las entradas por términos profesionales con ayuda de la base de datos SIDER, la cual contiene información de los medicamentos y las RAM que causan. Una vez con los datos preprocesados realizan una asociación entre los pacientes mencionados en las entradas y la información mencionada en las entradas del foro, después las entradas son analizadas para verificar si una entrada describe la experiencia de un paciente o no, tomando en cuenta el contexto de la oración, con esto obtuvieron cuatro variantes de su método propuesto que son PI (*Patient-centered and In-thread method*, Método Centrado en Paciente y en el Hilo) el cual solo contempla información del paciente en un mismo hilo, PA (*Patient-centered and Across-thread method*, Método Centrado en el Paciente a través de Hilos) el cual considera asociaciones entre hilos e información del paciente, PIE (*PI method with Experience mining*, Método PI con Minería de Experiencia) identificación de medicamento y RAM que fueron experimentados por un paciente en el mismo hilo y PAE (*PA method with Experience mining*, Método PA con Minería de Experiencia) identificación de medicamento y RAM que fueron experimentados por un paciente a través de varios hilos. El mejor método de identificación de RAM para las entradas evaluadas fue PI con una medida de F1 de 61.20%.

Vilar et al. (2018) realizaron una revisión sobre las metodologías y métodos de minería de datos para detectar DDIs (*Drug–drug interactions*, Interacción entre Medicamento-Medicamento), que es la coadministración de 2 o más medicamentos y que pueden afectar la salud de los pacientes, además los autores muestran que de las DDIs son responsables de causar aproximadamente el 30% de las RAM. Ellos dividieron los artículos en 3 enfoques que son: Farmacovigilancia, Literatura Científica y Medios Sociales. Centrándose en el enfoque de medios social, los autores analizaron 13 artículos de los cuales 5 se centran en estudios sobre la plataforma Twitter con una gran cantidad de volumen de tweets de aproximadamente 90 mil, estos estudios tuvieron resultados satisfactorios en extraer nombres de medicamentos, sustancias y DDIs. Así mismo otro estudio se centraba en Instagram fue un estudio realizado a lo largo de 5 años y recolectaron más de 5 millones de entradas relacionadas a medicamentos para tratar la depresión de más de 6 mil usuarios aportando así información de calidad a la farmacovigilancia, finalmente otros estudios se centraron en páginas web y foros tales como MedHelp. Por otra parte, los autores establecen que, si bien los trabajos existentes realizan aportes significativos, existen retos para aplicar minería de datos a la identificación de DDIs como lo son la falta de estandarización en nombres de medicamentos, la

disponibilidad de información ciertos medicamentos y el mismo uso del lenguaje natural por parte de las personas en medios sociales.

Gräßer et al. (2018) examinaron las opiniones de usuarios en el campo farmacéutico relacionado con la efectividad y efectos adversos de los medicamentos ya que el análisis de opiniones puede brindar puntos de vista valiosos y mejorar el monitoreo de la salud pública. Los autores recopilaron los datos de 2 páginas web Drugs.com y Druglib.com, páginas centradas en brindar información y proveer opiniones de medicamentos tanto de profesionales de la salud como personas comunes, utilizaron técnicas de *Web Scraping* con la librería *Beutiful Soup* del lenguaje de programación Python obteniendo el texto dentro de las etiquetas HTML y generaron 2 bases de datos la primera con 215063 opiniones obtenidas de Drugs.com y 3551 obtenidas de Druglib.com, y la segunda de nombre de medicamentos, 6345 obtenidos de Drugs.com y 541 obtenidos de Druglib.com. Para realizar la clasificación los autores proponen un método de regresión logística para clasificar los efectos adversos de los datos recopilados en la cual utilizaron los datos obtenidos de Druglib.com para el entrenamiento del modelo y como prueba del modelo utilizaron los datos de Drugs.com obteniendo una precisión del 49.75%, los autores concluyen que este resultado es debido a que los datos de entrenamiento no fueron suficientes para discriminar correctamente los datos de Drugs.com y en futuros trabajos proponen aumentar la cantidad de datos de entramiento.

Sarker et al. (2018) realizaron un sistema con el objetivo de evaluar los métodos de procesamiento automático de textos para la clasificación y normalización de contenidos relacionados a la salud en medios sociales, este sistema fue presentado en el congreso SMM4H (*Social Media Mining for Health*, Minería de medios sociales para la salud) en su edición de 2017. Los autores trabajaron entrenaron a su sistema con 15,717 tweets para la clasificación de RAM, 10,260 para identificar si realmente hubo un consumo del medicamento dividiéndolo en 3 clases (ingesta definida, posible ingesta y no ingesta) y 6,650 para la normalización de frases referentes a RAM. Así mismo, los autores evaluaron distintos métodos de clasificación divididos en 5 categorías: CNN, SVM, RNN, regresión logística y pilas de conjuntos; para la evaluación de la clasificación de RAM utilizaron 9,961 tweets obteniendo que el mejor método es un enfoque basado en SVM obteniendo un valor de F1 de 0.435, para la identificación del consumo de medicamentos el mejor método es un enfoque basado en una CNN que obtuvo un valor de F1 de 0.693, finalmente para la normalización de RAM se evaluaron 2,500 tweets con lo cual el método que se desempeñó mejor fue un enfoque basado en una RNN obteniendo una precisión del 88.7%, con lo cual los autores concluyen que los enfoques basados en SVM y redes

neuronales tienen un buen desempeño en estas tareas, sin embargo la clasificación de textos con lenguaje común sigue siendo una tarea retadora, dados los resultados de la clasificación de RAM.

Gupta et al. (2018) abordan el problema de la dependencia de datos etiquetados para la clasificación o extracción de datos, específicamente para la identificación de RAM, así mismo mencionan que el uso de redes neuronales profundas (*Deep neural networks*) suele presentar un costo computacional bastante alto debido al número de parámetros que deben ser ingresados y que dependen de datos etiquetados para su funcionamiento, sin embargo estos datos no siempre están disponibles, por lo cual los autores proponen un método semisupervisado basado en una RNN (*Recurrent Neuronal Network*, Red Neuronal Recurrente) para la extracción de RAM. Para la fase no supervisada utilizaron la RNN empleando una bolsa de palabras de medicamentos para tratar de predecir las menciones a medicamentos y reemplazarlos con un token denominado <*DRUG*> en textos de tweets, posteriormente para la fase supervisada reentrenaron la misma RNN utilizando un corpus de 645 tweets etiquetados y normalizados recuperados con la librería de Python tweepy, de los cuales utilizaron 470 para el entrenamiento y 170 para la evaluación, obteniendo un valor de F de 74.4% para esta evaluación y extracción de RAM, además realizaron experimentos con corpus obtenidos de *Google News* y documentos médicos, obteniendo un valor de F de 73.6% y 67.3% respectivamente, con estos resultados los autores concluyen que la estructura y semántica del lenguaje es lo más importante en las tareas de extracción de información aun cuando todos los corpus son del mismo dominio.

Suárez-Paniagua & Segura-Bedmar (2018) proponen un modelo basado en una CNN con el fin de extraer y clasificar DDIs evaluando distintas capas aplicables a la red neuronal de su modelo propuesto. Utilizan un corpus compuesto por datos anotados manualmente que contiene el nombre de 18,502 medicamentos y 5028 DDIs, extraídos de 1025 documentos médicos, así mismo los autores aplicaron un preprocesamiento convirtiendo todo el texto a minúsculas, removiendo caracteres especiales, remplazando números por el token *NUM* y finalmente remplazando el nombre de los medicamentos con los token *drugN,* estas oraciones normalizadas fueron convertidas en una matriz para ser la entrada del modelo propuesto ocupando 19,233 relaciones de DDIs y 4018 para la validación del modelo. Los autores realizaron 3 experimentos con capas diferentes para la CNN, estas capas fueron las siguientes: Máxima Agrupación, Agrupación Promedio y Agrupación Atenta, de las cuales la que mejores resultados obtuvo fue la capa de Máxima Agrupación al obtener un valor de F de 64.56% al evaluar el modelo. Así mismo los

autores mencionan que el combinar capas no presenta ninguna mejora en la clasificación demostrando que es mejor utilizar una capa de Máxima Agrupación para obtener los mejores resultados.

Identificar reacciones adversas causadas por medicamentos es una tarea muy importante y retadora debido que comúnmente estas reacciones no son reportadas formalmente, tal como lo menciona Chapman et al. (2019), es por lo que estos autores desarrollaron un sistema basado en NLP con el fin de identificar menciones de síntomas y medicamentos en notas clínicas y su relación entre una mención y un RAM. Los autores utilizaron un conjunto de datos de notas clínicas provisto por la Facultad de Medicina de la Universidad de Massachusetts, la cual está clasificada manualmente en las siguientes categorías: Medicamento, Indicación, RAM, Otras Señales, Frecuencia del Medicamento, Dosis, Duración del Medicamento, Vía de Administración y Severidad, además de incluir las relaciones entre Medicamento-Atributos, Medicamento-RAM, Medicamento-Indicación y Síntoma-Severidad. Para desarrollar la primera parte del sistema la cual se encarga de clasificar los datos en las clases mencionadas anteriormente los autores propusieron el uso de un algoritmo de aprendizaje máquina CRF (*Conditional random field*, Campo aleatorio condicional) debido a que es más rápido y menos demandante computacionalmente que un algoritmo de aprendizaje profundo; este algoritmo fue entrenado con distintos conjuntos de datos el primero con 100,000 notas clínicas de dominio público para realizar pruebas y encontrar la mejor combinación de parámetros, posteriormente se entrenó el algoritmo pero solo con 876 de los datos proporcionados por la universidad de Massachusetts, este algoritmo se evaluó con 213 notas clínicas obteniendo un valor de F de 80.9% para la clasificación. La segunda parte del sistema se encarga de identificar las relaciones existentes en las notas, para ello los autores utilizaron un modelo del algoritmo de bosque aleatorio (*random forest*) para evaluar este método utilizaron las 213 de la primera parte del sistema obteniendo un valor de F 88.1% en promedio para todas las relaciones existentes. Para la integración del sistema con ambos algoritmos en el estudio presentan un valor de F de 61.2% en lo cual los autores mencionan que se obtuvo un porcentaje menor de F debido a que los componentes no estaban correctamente ajustados para funcionar en conjunto.

En la investigación presentada por Wang et al. (2019) presentan un método para detectar potenciales RAM de manera automática mediante el uso de una red neuronal profunda. Los autores utilizaron datos de SIDER el cual es un repositorio de información de medicamentos, tomaron datos de los años 2009 y 2012, obteniendo 978 nombres de medicamentos y 1325 efectos adversos relacionados a

estos, además extrajeron descriptores químicos y propiedades biológicas que servirán para brindar más información al modelo desarrollado. Como primera parte del entrenamiento de la red neuronal profunda utilizaron 2.3 millones de artículos biomédicos con el fin de crear un vector denominado D2V para que el modelo pudiera entender las características semánticas de medicamentos de los cuales solo obtuvieron datos para 764 medicamentos. El modelo desarrollado utilizó como entrada el vector D2V y se comparó frente a 2 métodos habituales de aprendizaje profundo que son un clasificador lineal de vectores de soporte y un clasificador Gaussiano Naïve-Bayes, este experimento se realizó tratando de predecir RAM e identificando RAM, obteniendo para el método propuesto un promedio de precisión para la predicción de RAM de 72.1%, así como para la identificación un 38.2% y la combinación de ambos un 52.3%, desempeñándose mejor que los métodos habituales siendo que el clasificador lineal obtuvo un 17% en predicción de RAM, 2.5% en la identificación y de manera general un 19.5% de precisión, mientras que el clasificador Naïve-Bayes obtuvo un 34% en la predicción de AD, un 6% en la identificación y de manera general un 9% de precisión. Los autores mencionan que el motivo principal por el cual la capacidad de identificación del modelo no es tan alta es porque usualmente los RAM no son reportados formalmente en los datos lo cual hace de esta tarea algo especialmente retador, sin embargo, el modelo demuestra gran capacidad de predicción de RAM incluso para nuevos medicamentos lo cual vuelve al modelo desarrollado una gran herramienta para la identificación temprana de RAM.

Alimova & Tutubalina (2020) presentan un análisis de efectividad sobre los modelos de redes neuronales más utilizados para la clasificación de RAM contrastados con uno de los mejores modelos existentes basado en una SVM. Las autoras han encontrado que la mayoría de los métodos solo se prueban con un conjunto de datos, es por lo que ellas proponen 4 corpus diferentes para realizar los experimentos, los cuales se describen a continuación, 1) CADEC, corpus compuesto de comentarios etiquetados en medicamento, efecto adverso, enfermedad, síntoma y otro, obtenidos de usuarios del foro askapatient.com; 2) Twitter, consiste en tweets relacionados a la salud donde las entidades reacción adversa y enfermedad se encuentran marcadas; 3) MADE, compuesto registros médicos de pacientes con cáncer y 4) Twimed, compuesto de tweets y artículos obtenidos de PudMed, con anotaciones de enfermedad, síntoma y medicamento. Las arquitecturas de redes neuronales utilizadas en la comparación fueron las siguientes, 1) LSTM, es un tipo de RNN que utiliza como entrada un vector de palabras el cual es pasado a la siguiente capa y guarda el resultado en estados ocultos, los cuales después de leer toda la oración palabra por palabra son enviados a la capa de clasificación con una

función conocida como softmax regresando el RAM etiquetado; 2) TD-LSTM, esta arquitectura es una extensión de la anterior tomando en cuenta el contexto tanto del lado izquierdo como derecho de la oración concatenando los resultados de los estados para generar un resultado; 3) IAN, modelo atención interactiva que utiliza capas LSTM para producir vectores que calculan la atención del vector de palabras los cuales posteriormente se concatenan y se utiliza una capa softmax para producir el resultado; 4) RAM, el modelo procesa el contexto utilizando capas bidireccionales de LSTM produciendo vectores de información y guardándolos en memoria, después se identifica la entidad a ser clasificada, utilizando una capa bidireccional LSTM produciendo vectores con estados ocultos y la tercera aplica mecanismos de atención a los vectores 1 y 2 para posteriormente concatenar los resultados y aplicar una capa softmax y 5) MemNet, modelo compuesto por un módulo de memoria que guarda el contexto de la oración y una capa de atención en la cual se generan vectores que se van sumando y pasan a la siguiente capa hasta producir una salida. Finalmente, para la clasificación de RAM el mejor modelo fue IAN ya que obtuvo los mejores resultados en términos del valor de F en 3 de 4 corpus (81.5% para corpus 1, 78.6% para corpus 3 y 84.65% para corpus 4), mientras que el modelo RAM se desempeñó mejor en el corpus 2 de Twitter obteniendo un valor de F de 83.4%, concluyendo así que los modelos basados en RNN son igual de eficaces que un modelo basado en el método SVM e incluso desempeñándose mejor las RNN que tienen memoria adicional y módulos de atención.

Edo-Osagie et al. (2020) realizaron una revisión sobre el impacto actual de Twitter como instrumento para los estudios de la salud pública y para identificar los principales campos en los que se desarrollan estos estudios. Los autores analizaron 92 artículos en idioma inglés de un periodo entre 2009 a 2019 los cuales cumplían con sus criterios de inclusión, en estos encontraron lo siguiente: 82 artículos se centran en problemas de salud de los cuales el tema más común son las enfermedades, seguido de abuso de medicamentos y efectos adversos causados por medicamentos, entre temas menos investigados se encuentran la farmacovigilancia, seguimiento de enfermedades y prevención. Así mismo las técnicas más utilizadas incluyen métodos de aprendizaje supervisado tales como SVM o regresión logística con 70 artículos, por otra parte el aprendizaje no supervisado y semisupervisado se utiliza en 18 y 4 estudios, respectivamente, y el aprendizaje profundo se utiliza en 16 artículos, los autores mencionan que el motivo por el cual el método SVM es tan utilizado es por su popularidad y su robustes ante problemas de clasificación de texto, sin embargo recalcan que a partir del año 2018, lo métodos basados en aprendizaje profundo como redes neuronales se han hecho más populares y marcan una tendencia dominante para el análisis de datos obtenidos de Twitter. Para la

detección de RAM los autores comprobaron que el 40% de los estudios abordan esta problemática, siendo que las investigaciones son bastante recientes la mayoría presentándose desde el 2016, además en estos estudios la técnica para la identificación de RAM más común es el aprendizaje supervisado utilizando un método basado en SVM. Como conclusión los autores señalan que la aplicación de datos de Twitter enfocado en la salud publica aún tiene retos interesantes no cubiertos a pesar de los avances significativos que se han realizado en los últimos años, así mismo este análisis ayuda a identificar huecos que falta cubrir en el estado del arte permitiendo a los investigadores enfocar sus estudios de manera específica a los problemas existentes, finalmente con este análisis se demuestra que el uso de Twitter ha dado resultados positivos en distintos dominios como los mencionados anteriormente.

Las reseñas escritas por pacientes acerca de medicamentos son muy importantes por el contenido que tienen, ya que ofrecen datos médicos sobre el paciente, su tratamiento y satisfacción o frustración por el servicio sanitario, en este sentido Basiri et al. (2020) proponen 2 modelos basados en la teoría de decisión a tres bandas para analizar las reseñas de medicamentos. Los autores utilizaron un conjunto de datos de 215,063 reseñas de medicamentos obtenidas de Drugs.com y categorizadas en positivas, negativas y neutrales, así mismo aplicaron un preprocesamiento para obtener las reseñas que contenían entre 85 y 180 palabras que representaban el promedio de tamaño de las reseñas quedándose con 214,600, posteriormente separaron cada reseña en una lista de palabras, definieron palabras clave y separaron el conjunto de datos en datos para entrenamiento y de prueba. Para el primer modelo propuesto los autores utilizaron una fusión del método de aprendizaje máquina y método tradicional este modelo es llamado 3W1DT, en este modelo ambos ejecutan la clasificación y si el método de aprendizaje máquina produce un valor de clasificación de alta confiabilidad de acuerdo con un límite definido se toma su resultado, en caso contrario se toma el valor método tradicional. Para el segundo modelo denominado 3W3DT se utilizan 3 métodos de aprendizaje profundo y un método tradicional, para obtener un resultado del modelo se comparan los resultados obtenidos por cada método y se da como valido aquel que tenga el valor máximo de la clasificación. Los autores demostraron que para 3W1DT los mejores métodos a utilizar es un clasificador 3CRNN (*Three-way Convolutional Recurrent Neural Network*, Red Neuronal Convolucional Recurrente a 3 Bandas) combinado con el método Naïve-Bayes obteniendo un valor de F de 87%, mientras que para el modelo 3W3DT lo mejores métodos son GRU (*Gated Recurrent Unit*, Unidad Recurrente Cerrada), CNN y 3CRNN combinados con el método Naïve-Bayes obteniendo un valor de F de 87.35%. Como conclusión en el estudio se menciona

que el mejor modelo es 3W3DT al obtener los mejores resultados, además de haber sido comparado contra otros métodos del estado del arte actual como lo son AC-BiLSTM, IWV (*Improved Word Vectors*, Vectores de Palabras Mejorados), CRNN y ARC (Attention Neural Network, Red Neuronal de Atención) y obteniendo resultados superiores por hasta un 4%, además los autores mencionan que el modelo puede ser utilizado en otros problemas de clasificación siempre y cuando el entrenamiento se adapte al problema específico.

Lavertu et al. (2021) desarrollaron un método semisupervisado para estimar el grado de gravedad de RAM utilizando incrustaciones de palabras extraídas a partir de redes sociales. Para la preparación de los datos utilizaron el diccionario MedDRA para extraer RAM y sus sinónimos, posteriormente para el modelo de incrustaciones de palabras los autores lo obtuvieron de RedMed, el cual es un modelo enfocado en términos médicos preprocesados donde se maximiza la inclusión de RAM extraídos de la red social Reddit en el cual se incluían 28,113 RAM, además se incluyó la clasificación de severidad de RAM de acuerdo con el sistema FAERS (*Food and Drug Administration Adverse Event Reporting System*, Sistema de notificación de acontecimientos adversos de la Administración de Alimentos) que incluye las clases de ”muerte”, “resultado grave” y “sin resultado”. Los autores construyeron una red para la estimación de severidad la cual fue entrenada con el modelo de incrustación de palabras utilizando el método de caminos aleatorios, obteniendo los siguientes resultados de acuerdo con el índice de correlación de Spearman, un 0.595 para RAM que tenían correlación con la clase de “muerte”, un 0.633 para la clase de “resultado grave” y 0.748 para “sin relación”. Finalmente, los autores concluyen que el uso de redes léxicas y propagación de etiquetas de su modelo propuesto puede utilizarse para estimar de manera cuantitativa la gravedad de las afecciones médicas, además también mencionan que las reacciones adversas que tienen una mayor correlación con la clase “muerte” se da durante el periodo de comercialización y son significativamente mayores que las descubiertas durante ensayos clínicos.

Dolores Barrientos et al. (2022) en el capítulo 14 de su libro presentan la dimensión actual sobre como las noticias falsas que se encuentran en las redes sociales afectan a las personas por consumir medicamentos o productos milagro, los cuales son recomendados por personas que no tienen experiencia en el ámbito de la salud y que por la necesidad que tienen las personas de encontrar una cura a enfermedades tales como el COVID-19 los ingieren sin conocer los riesgos a su salud. En su análisis los autores proponen 3 puntos importantes, el primero acerca de las noticias falsas de medicamentos de patente, estos son medicamentos aprobados por la COFEPRIS sin embargo, en redes como Facebook y YouTube se compartieron tratamientos con

medicamentos para enfermedades como el COVID-19, tal es el caso del medicamento Esteriflu, Esticide, Hidroxicloroquina e Ivermectina, los cuales eran promocionados con este fin siendo que no había estudios acerca de la efectividad de estos medicamentos para tratar COVID-19 desinformando a la población; el segundo sobre noticias falsas de productos milagro, estos productos aumentaron su comercialización después de la pandemia sin embargo, estos no son avalados por la COFEPRIS y presentan un riesgo para la salud de la población, además hacen uso de las emociones y sentimientos de la población con mensajes publicitarios altamente persuasivos para su compra y consumo; finalmente el tercer aspecto son las noticias falsas de remedios caseros, estas suelen compartirse entre familiares y amigos por grupos de WhatsApp, promoviendo bebidas, nebulizaciones, consumo de hierbas, etc., provocando en muchas ocasiones que las enfermedades avancen y los pacientes lleguen en estado grave a los hospitales. Con el análisis realizado los autores concluyen que gracias a la pandemia se ha puesto en evidencia la facilidad con la cual las noticias falsas acerca de medicamentos, productos milagro y tratamientos caseros son compartidos a través de redes sociales y como estas mismas juegan un rol muy importante debido a la velocidad con la cual se puede compartir una noticia falsa, por lo cual los autores invitan a la reflexión social y a como se trata la información y así evitar la infodemia.

Sakhovskiy & Tutubalina (2022) realizaron un modelo basado en BERT (*Bidirectional Encoder Representations from Transformers*, Representaciones de codificadores bidireccionales a partir de transformadores) para la clasificación de tweets con el fin de identificar RAM utilizando un enfoque intuitivo donde mencionan que el texto y la estructura de los medicamentos son complementarios. En el estudio utilizaron 3 corpus diferentes proporcionados por SMM4H en su edición de 2020, donde proporcionaron tweets en inglés (18299), francés (3033) y ruso (20707) anotados manualmente en la cual se incluía información sobre la presencia, ausencia o mención específica de RAM. Para preprocesar el texto de los tweets utilizaron 3 versiones de BERT cada una dependiendo del idioma RoBERTa para los tweets en inglés, EnRuDR-BERT para los tweets en ruso y CamemBERT para los tweets en francés, de los medicamentos obtuvieron datos bioinformáticos y químicos, además de su ATC (*Anatomical Therapeutic Chemical*, Química Anatómica Terapéutica), y finalmente sus datos moleculares obtenidos de la librería Mordred. Finalmente, para la clasificación utilizaron una red neuronal conectada completamente incluyendo una capa oculta, activaciones GeLU y una salida de 0.3%, además utilizaron una capa sigmoidea como activación de la capa de salida y entropía cruzada binaria como función de perdida. Esta red fue comparada contra 3 modelos una SVM, una CNN y un clasificador BERT clásico, obteniendo los siguientes resultados: para la

clasificación de RAM en tweets en francés se obtuvo una mejora en los resultados del 8.5% en términos de F1 con respecto al estado del arte donde la mejor combinación fue BERT-MolBERT, para los tweets en ruso el mejor modelo fue igualmente la combinación de BERT-MolBERT obteniendo una mejora del 0.3% en términos de F1 con respecto al estado del arte y por último, para los tweets en inglés donde el mejor modelo fue el que incluyó RoBERTa y descriptores para la clasificación obteniendo un 18% de mejora en términos de F1 con respecto al estado del arte. Los autores concluyen que utilizar modelos BERT es de gran utilidad, al ser modelos bien conocidos y dando la posibilidad de mejorar el estado del arte actual, además como trabajos a futuro mencionan la posibilidad de incluir conjuntos de datos en múltiples lenguajes y el aprendizaje para múltiples tareas.

2.2 Análisis comparativo de los trabajos relacionados

De acuerdo con la Tabla 1, se presenta la comparación de los trabajos relacionados que fueron analizados en el apartado anterior, presentando datos más relevantes de los estudios, con el fin de evidenciar las diferencias y similitudes de las contribuciones existentes, las tecnologías utilizadas y los resultados obtenidos, contrastarlos así contra los datos del proyecto desarrollado.

Tabla 1 Comparación entre trabajos existentes y sus principales características.

Autor	**Enfoque / Problemática**	**Tecnologías utilizadas**	**Arquitectura**	**Resultados**
Zunic et al. (2020)	Revisión sistemática sobre el estado del arte, técnicas utilizadas y problemas abordados sobre el análisis de sentimientos en el ámbito de la salud y bienestar.	-	No	Identificación de 86 artículos relevantes e identificación de 5 roles desde las perspectivas de Enfermo, Adicto, Paciente, Cuidador y Suicida.
Liu et al. (2018)	Identificación de efectos adversos de medicamentos centrado en la perspectiva del paciente utilizando opiniones del foro de enfermedades cardiacas MedHelp.	Base de datos SIDER, Núcleo de herramientas de Standford y Meta-Mapeos del Sistema de Lenguaje Médico Unificado (UMLS), Reconocimiento de Entidades Nombradas (NER), Parte del Discurso (POS)	Si	Técnica denominada "Método Centrado en Paciente y en el Hilo" con una medida de F1 de 61.20% para la identificación de RAM.
Vilar et al. (2018)	Analizar los trabajos relacionados a DDIs, así mismo se analiza el enfoque	-	No	Análisis de 13 artículos identificando 3 roles Farmacovigilancia, Literatura Científica y Medios Sociales, establecimiento retos para NLP,

	de minería de datos en la literatura existente y su uso en medios sociales.			búsqueda de medicamentos y medios sociales.
Gräßer et al. (2018)	Examinar opiniones de usuarios sobre el campo farmacéutico acerca de satisfacción y efectos adversos de medicamentos utilizando aprendizaje máquina y datos de las páginas web Drugs.com	Base de datos obtenida de Drugs.com, aprendizaje máquina (regresión lógica mediante unigramas, bigramas y trigramas) y Web Scraping (utilizando beautiful soup).	Si	Desarrollo de modelo utilizando aprendizaje máquina y obteniendo una precisión del 49.75%
Sarker et al. (2018)	Desarrollar métodos de procesamiento automático de texto para la clasificación y normalización de tweets relacionados a la salud.	Twitter API, SVM, CNN, RNN, regresión logística y pilas de conjuntos.	Si	En el estudio se realizaron pruebas sobre los 5 métodos desarrollados obteniendo un valor de F1 de 0.435 para la clasificación de RAM, y los autores concluyen que la clasificación de textos de lenguaje natural sigue siendo una tarea retadora.
Gupta et al. (2018)	Desarrollo de método para la extracción de RAM mencionados en tweets, y disminuir la necesidad de utilizar datos etiquetados.	Twitter API, Tweepy, RNN bidireccional, Keras, word2vec.	Si	En el estudio obtuvieron un valor de F de 74.4% para la extracción de RAM con tweets, un valor de F de 73.6% para Google News y un valor de F de 67.3% para documentos médicos, los autores concluyen que los más importante es la estructura y semántica de los corpus.

Suárez-Paniagua & Segura-Bedmar (2018)	Extracción de DDIs mediante el uso de una red neuronal para obtener el mejor resultado de la extracción en documentos médicos.	Base de datos DDI-Medline, herramienta Brat rapid, CNN.	Si	En el estudio desarrollaron una CNN en la cual probaron 3 capas distintas, con la que se obtuvo un mejor resultado fue la de Máximo Agrupamiento obteniendo un valor de F de 64.56%
Chapman et al. (2019)	Identificar RAM no reportados debido a que es una tarea muy importante y que presenta diversos retos.	Base de datos de UMASS, NER, RE, CRF.	Si	En el estudio se evaluaron 2 partes de un sistema en la detección de RAM se obtuvo un valor de F de 80.9% y en la identificación de relaciones un valor de F de 88.1%, para el sistema combinado se obtuvo un valor de F 61.2% donde los autores indican que el valor disminuyó debido a una falta de ajuste en el funcionamiento conjunto de las partes.
Wang et al. (2019)	Identificar RAM no reportados, predecir RAM de nuevos medicamentos analizando información biomédica, química y biológica para mejorar la detección de estos.	Bases de datos SIDER, PubChem DrugBank y MEDLINE, Modelo basado en una DNN,	Si	Para el modelo propuesto se obtuvo un 72.1% en la predicción de RAM, un 38.2% para la identificación y un 52.3% para la combinación de ambas partes, este modelo demuestra una gran capacidad de predicción.
Alimova & Tutubalina (2020)	Análisis de efectividad sobre los modelos neuronales para la	Corpus: CADEC, Twitter, MADE y Twimed	Si	En el estudio se comprobó que el mejor método es IAN al obtener los mejores resultados 3 de 4 corpus y el modelo RAM obtuvo el mejor resultado para el corpus

	identificación de RAM y validación contra modelo SVM.	Redes Neuronales: LSTM, TD-LSTM, IAN, RAM. Algoritmo de aprendizaje máquina: SVM		de Twitter obteniendo un valor de F de 83.4%
Edo-Osagie et al. (2020)	Revisión sobre el impacto actual de Twitter como instrumento para los estudios de la salud pública e identificar los principales campos en los que se desarrollan estos estudios.	-	No	Analizaron 92 artículos donde el 40% de estos se centran en la problemática de los RAM, así mismo los autores concluyen que el uso de Twitter ha dado resultados positivos enfocados en la salud publica presentando aún retos a pesar del avance tecnológico.
Basiri et al. (2020)	Análisis de reseñas escritas por pacientes con la finalidad de procesarlas y obtener información relevante sobre opiniones de medicamentos.	Base de datos de Drug.com, CNN, Naïve-Bayes, árbol de decisión, GRU, 3CRNN	Si	En el estudio se desarrollaron 2 modelos denominados 3W1DT y 3W3DT obteniendo un valor de F de 87% y 87.35% para la clasificación de reseñas, respectivamente. En el estudio se menciona que estos modelos pueden ser utilizados para otros contextos con el entrenamiento adecuado.
Lavertu et al. (2021)	Surge de la necesidad de estimar el grado de gravedad de RAM apoyándose de la información existente en redes sociales.	MedDRA, Reddit, incrustación de palabras, caminos aleatorios.	Si	En el estudio obtuvieron la correlación de Spearman entre RAM y las clases "muerte" con 0.633, "resultado grave" con 0.748 y "sin resultado" con 0.595,

				demostrando que el modelo puede utilizarse para estimar correctamente la gravedad de los RAM.
Dolores Barrientos et al. (2022)	Analizar los efectos de las noticias falsas en redes sociales principalmente relacionadas con el uso de fármacos y productos milagro y su efecto en la salud de las personas.	-	No	En el estudio se concluye que después de la pandemia se acentuó la facilidad de transmitir noticias falsas a través de redes sociales, además de que esto puede dañar a la salud de las personas por consumir medicamentos o productos milagro no avalados por entes como la COFEPRIS.
Sakhovski y & Tutubalina (2022)	Clasificación de tweets como fuentes de RAM incluyendo la relación con los medicamentos y su estructura molecular.	Clasificadores unimodales BERT (RoBERTa, MolBERT, ChemBERTa), SVM, CNN, Drugbank	Si	Se puede observar una mejora del estado del arte actual en términos de F1 en tres corpus en diferentes idiomas, para tweets en francés un 8.5%, para tweets en ruso una mejora del 0.3% y para los tweets en inglés un 18%.

Después de realizar un análisis exhaustivo sobre el estado del arte actual, se encontró que la identificación y clasificación de RAM es un problema ampliamente estudiado, así mismo en 8 de los estudios analizados hacen uso de comentarios en redes sociales como entrada de datos para la clasificación de RAM haciendo importante el análisis de los comentarios al brindar información relevante sobre el consumo de medicamentos para un conjunto diverso de enfermedades tales como la diabetes, el cáncer, la hipertensión o enfermedades comunes como la gripe. Así mismo estos artículos utilizan técnicas y modelos de aprendizaje máquina, aprendizaje supervisado y semisupervisado y aprendizaje profundo siendo estos últimos los que han presentado mejores resultados en los estudios más recientes.

Algunos retos y áreas de oportunidad que se identificaron en estos estudios es la gran cantidad de datos que toma entrenar los modelos para obtener resultados con un alto grado de precisión, otro punto a mejorar es la entrada de datos tal como lo menciona Gupta et al. (2018) la semántica de los comentarios y registros médicos es de vital importancia ya que un modelo puede estar ajustado para identificar cierto significado dependiendo del tipo de textos lo cual puede modificar el grado de precisión en la identificación de RAM. Así mismo estos modelos suele ajustarse a un idioma en particular y aunque como lo menciona Basiri et al. (2020) los modelos pueden reentrenarse para realizar la identificación de RAM esto no se ha llevado a cabo en la práctica, tal y como se evidencia en Sakhovskiy & Tutubalina (2022) donde utilizaron tweets en francés y ruso que fueron presentados por primera vez por SMM4H en el año 2021.

De acuerdo con lo anterior, se hace evidente la importancia de desarrollar esta propuesta, teniendo un enfoque en el idioma español y así aportar al desarrollo del estado del arte, abordar el problema de los RAM en México y aprovechar las opiniones de los usuarios de redes sociales como fuente de información.

2.3 Alternativas de solución

2.3.1 Capa de datos

A continuación, se listan una serie de bases de datos con el fin de contrastar sus características y poder elegir la mejor para la propuesta de solución.

Bases de datos SQL

Tabla 2 Comparativa de características entre principales bases de datos SQL en la actualidad.

Nombre	Características clave	Uso Principal
SQL Server	• Soporte de transacciones, escalabilidad, estabilidad y seguridad. • Extensión del lenguaje SQL mediante Transact-SQL. • Licencia de tipo propietario (desde 899 USD anuales) para poder utilizar la mayoría de sus características. • Almacenamiento en la nube para las licencias *Enterprise*. • Permite trabajar bajo esquema cliente-servidor.	Bases de datos principalmente montadas en entornos de Windows Server, aunque en sus últimas versiones puede ser montada en Red Hat, Ubuntu y SUSE.
MySQL	• Compatibilidad completa con el lenguaje estándar SQL. • Desde su versión 5.0 da la posibilidad de crear vistas para tablas de gran tamaño. • Manejo de transacciones y procedimientos almacenados. • Implementación de desencadenantes para automatizar tareas. • Es una BD de código abierto por lo cual no es necesario adquirir una licencia.	Puede ser montado en cualquier sistema operativo, siendo compatible principalmente con sistemas basados en Linux, y aplicable principalmente a pequeñas y medianas empresas al ser de código abierto.

PostgreSQL	• Enfocado en la alta concurrencia permite atender a varios clientes al mismo tiempo. • Soporte de distintos tipos de datos de manera nativa, tales como direcciones IP, direcciones MAC, arreglos, figuras geométricas, entre otros. • Permite utilizar un enfoque basado en Objeto-relacional, un ejemplo es la herencia de tablas. • Funciones de seguridad e integridad referencial.	Al igual que MySQL puede ser mondado en cualquier sistema operativo al ser de código abierto. Principalmente utilizado gracias a su robustes y características, puede ser usado desde pequeñas hasta grandes empresas.

Bases de datos No SQL

Tabla 3 Comparativa de características entre principales bases de datos no SQL en la actualidad.

Nombre	**Características clave**	**Uso Principal**
Firebase	• Almacena y sincroniza datos en tiempo real, útil para aplicaciones como chats o que necesitan ver reflejados los cambios de manera automática. • Permite uso sin conexión, utilizando la caché del programa y posteriormente sincronizando los datos. • Establecimiento de normas de seguridad. • Guardado de los datos en documentos o colecciones en formato no estructurado JSON. • Confiabilidad, respuesta rápida y gran manejo de volumen de datos. • Permite un uso de prueba gratuito sin embargo a partir de 1GB de almacenamiento cobra 0.90 USD por cada 100,000 datos almacenados.	Aplicaciones que requieren actualización de manera automática, con implementación de manejo de datos preconstruida y almacenamiento en la nube, tomando en consideración que si la aplicación crece de manera exponencial el costo puede hacerlo de igual manera.

MongoDB	• BD no relacional que permite el guardado de documentos JSON en colecciones e incluso manejo de archivos. • Permite consultas por campos, por rangos de campos y expresiones regulares. • Permite crear índices de todos los campos y subíndices mejorando el tiempo de respuesta al realizar consultas. • Es posible crear un balanceador de carga que permite escalar y mantener el servicio operativo ante alguna falla de un servidor.	Al ser de código abierto y escrito en C++ es posible montarlo en distintos servidores aprovechando el potencial de estos, así mismo permite el manejo de un gran volumen de datos y además el almacenado de documentos es muy flexible.
DynamoDB	• BD no relacional, pero gracias a PartiQL es compatible con consultas SQL. • Es la que mejor maneja grandes cantidades de datos permitiendo almacenar/consultar hasta 10 billones de datos al día y lecturas de 20 millones de datos por segundo. • Permite la administración automática, con escalado automático y modo de lectura y escritura bajo demanda. • Al ser una BD propietaria se debe pagar para su uso, este se calcula por documentos por segundo, teniendo un costo escritura de $0.18 USD por 12 unidades por segundo y $0.03 USD por 12 unidades por segundo para lectura.	Aplicable principalmente a grandes organizaciones dado el costo de uso, sin embargo, ofrece grandes beneficios al administrar grandes volúmenes de datos.

2.3.2 Capa de presentación

A continuación, se listan los principales marcos de trabajo para el desarrollo de capas de presentación, la mayoría de estos marcos de trabajo integran las tecnologías de JavaScript, HTML5 y CSS3.

Tabla 4 Comparativa de tecnologías para desarrollo de capa de presentación.

Nombre	**Características clave**	**Tipo de aplicación**

React y React Native	• Desarrollo en base a la creación de componentes. • Flujo de datos unidireccional. • Creación de aplicaciones de alta demanda. • Desarrollo declarativo, basado en estados que son actualizados a lo largo de la aplicación. • Es isomórfico, lo cual le permite ejecutar código del lado del cliente y del servidor.	React permite la creación de aplicaciones SPA (*Single Page Application*, Aplicaciones de Página Única), mientras que React Native permite la creación de aplicaciones móviles nativas para Android y iOS.
VueJS	• Es un *framework* estándar y que implementa las tecnologías base para la creación de aplicaciones web. • Desarrollo en base a componentes. • Sigue un patrón determinado por HTML, en la que primero se encuentran etiquetas, después funciones y métodos y al final estilos.	Creación de aplicaciones SPA reactivas, manejando todo del lado del cliente y teniendo una aplicación ligera y modular.

2.3.3 Capa de dominio

A continuación, se listan los principales marcos de trabajo para el desarrollo de capas de presentación, la mayoría de estos marcos de trabajo integran las tecnologías de JavaScript, HTML5 y CSS3.

Tabla 5 Comparativa de características de tecnologías para desarrollo de capa de dominio.

Nombre	Características clave	Tipo de aplicación
NodeJS	• Es un *framework* de capa de servidor construido sobre el motor de JavaScript 8 de Google Chrome lo cual lo hace veloz. • No maneja búfer ya que genera sus datos en manera de chunks.	Diseñado principalmente para realizar aplicaciones escalables y gestión de

	• Es asíncrono y controlado por eventos en donde no se espera que terminen las llamadas a funciones, sino que se notifican por medio de eventos. • Modelo basado en un subproceso cíclico que responde sin bloqueos lo cual lo hace altamente escalable.	múltiples conexiones internas y externas al mismo tiempo.
Laravel	• Es un *framework* de PHP basado en la arquitectura MVC (Modelo-Vista-Controlador) lo cual lo hace modular y permite un crecimiento robusto y estable según el sistema vaya creciendo. • Cuenta con un potente ORM (mapeo objeto-relacional), el cual es compatible con bases de datos relacionales y no relacionales. • Posibilidad de utilizar librerías de terceros. • Para la manipulación de datos soporta los modos de trabajo *code first* (primero código) o *database first* (primero base de datos).	Creación de aplicaciones Web en su completo o la posibilidad de crear APIs REST que pueden ser expuestas, así mismo soporta autenticación JWT que permite el escalamiento de las aplicaciones mediante un balanceador de carga.

2.4 Solución propuesta

Tabla 6 Tecnologías elegidas para el desarrollo del sistema de identificación de efectos adversos.

Base de datos	MongoDB
Lenguajes de Programación	PHP, Python, JavaScript, TypeScript
Metodología de desarrollo	SCRUM

Frameworks de desarrollo	Laravel (Backend), React Native (Frontend), VueJS (Frontend)
APIs	API de Twitter
Método de procesamiento de lenguaje natural	RAM (Red Neuronal de Atención Recurrente)

2.4.1 Justificación de la solución seleccionada

Para la capa de datos se decidió utilizar MongoDB debido a las características con las que cuenta, al ser una base de datos no relacional y almacenar los datos en documentos en formato JSON, el cual es la base de las respuestas de APIs como la que ofrece Twitter y debido a su uso flexible y gran manejo de volumen de datos esta es la mejor opción para el desarrollo del proyecto. Así mismo los lenguajes de programación a utilizar dependen de los *frameworks* elegidos para desarrollar la capa de dominio y la capa de presentación, en la capa de dominio se eligió Laravel debido a su arquitectura interna basada en MVC, al contar con un ORM potente para consultas relacionales y no relacionales y al poder incluir autenticación pasada en JWT lo cual hace que las APIs desarrolladas sean seguras y solo las aplicaciones que tengan el permiso puedan conectarse a la capa de dominio. Por la parte de la capa de presentación se utilizará React y React Native dada su gran capacidad de adaptación, el desarrollo por módulos y la posibilidad de crear aplicaciones SPA y una aplicación móvil.

Para el análisis de datos de acuerdo con el análisis realizado en el estado del arte se puede apreciar que un modelo basado en una red neuronal de atención recurrente es de los mejores modelos para la identificación de efectos adversos en medicamentos, así mismo el análisis y la programación de este modelo se llevará a cabo con Python debido a las herramientas disponibles actualmente.

CAPITULO III METODOLOGÍA Y DESARROLLO

En este capítulo se presentan las metodologías aplicadas para cumplimiento del objetivo de esta investigación. Para la obtención de datos se utiliza la metodología CRISP-DM describiendo la aplicación de sus distintas fases, mientras que para el desarrollo del sistema Web y Aplicación Móvil se ha uso de la metodología SCRUM.

3.1 Metodología CRISP-DM

Dada la necesidad de trabajar con gran volumen de datos, a la flexibilidad y adaptabilidad que brinda la metodología CRIPS-DM, se hace uso de la misma para el desarrollo del modelo de datos. A continuación, se describirá la aplicación cada una de las seis fases del ciclo de vida de esta metodología para la creación del modelo de datos de este proyecto.

3.1.1 Entendimiento del Negocio

Para la realización de este proyecto se identifica la necesidad de recopilar información relacionada con la toma de medicamentos y las reacciones adversas no reportadas que causan a las personas. Sin embargo, debido al gran número de diferentes enfermedades, a la gran cantidad de medicamentos y a los distintos efectos adversos existentes, es importante delimitar la recopilación de información con el fin de obtener datos más precisos y enfocados en padecimientos específicos.

Las enfermedades consideradas para la búsqueda de información son la diabetes mellitus tipo 1 y 2, así como la hipertensión. Estas enfermedades fueron seleccionadas, ya que acuerdo con el INEGI en el 2018 en México existían 8,542,718 personas diagnosticadas con diabetes, además para 2020 la diabetes pasó a ser la tercera causa de muerte en el país. Por su parte la hipertensión afecta a más de 30 millones de personas de acuerdo con el INEGI (2020) y ocasiona cerca de 50 mil fallecimientos cada año. De acuerdo con estos datos, se hace evidente que estas enfermedades tienen una gran presencia en México ya que afectan a casi el 30% de la población. Otro punto importante para este proyecto es la recolección de información sobre alertas sanitarias emitidas por la COFEPRIS, ya que estas alertas pueden brindar información de gran utilidad a la población y normalmente no son visitas por un gran número de personas.

Con base en lo anterior se desprenden los siguientes objetivos del negocio del modelo:

- Construir un corpus de comentarios de personas para la identificación de efectos adversos causados por medicamentos.
- Recolectar información de medicamentos y efectos adversos y sus características.
- Desarrollar módulo de obtención de alertas sanitarias emitidas por la COFEPRIS.
- Clasificar comentarios de acuerdo con la presencia o ausencia de la mención de un efecto adverso causado por medicamentos.

Plan de desarrollo del modelo

Para el cumplimiento de los objetivos del negocio, se seguirán las siguientes fases:

- Fase 1: Diseñar módulo de recolección de comentarios, diseño y construcción de base de datos para medicamentos y efectos adversos, diseño de módulo de obtención de alertas sanitarias.
- Fase 2: Desarrollo de módulo de recolección de comentarios y desarrollo de módulo de obtención de alertas sanitarias.
- Fase 3: Preparación de los datos obtenidos (formateo y limpieza para los datos).
- Fase 4: Elección de modelos de PLN para clasificación de comentarios.
- Fase 5: Análisis de resultados obtenidos de la etapa anterior.
- Fase 6: Presentación de resultados.

3.1.2 Entendimiento de los Datos

En esta segunda etapa de la metodología CRIPS-DM es necesario realizar la recolección de los datos iniciales del modelo con el fin de conocer sus relaciones y que estructura pueden tomar a partir de ello.

Recolección de datos iniciales

Antes de realizar la recolección de datos es importante conocer el qué buscar y el porqué de esa decisión, en el apartado anterior se mencionó que las enfermedades consideradas son la hipertensión y diabetes mellitus tipo 1 y tipo 2, es por ello que buscar medicamentos relacionados con estas enfermedades es de gran relevancia, por lo que se realizó una consulta a una doctora quien proporcionó los nombres de los medicamentos más utilizados para tratar estas enfermedades tal y como se muestra en la Tabla 7.

Tabla 7 Conjunto de medicamentos para el tratamiento de diabetes e hipertensión.

Medicamentos Diabetes	Medicamentos Hipertensión

Metformina	Semaglutida	Liraglutida	Losartan	Quinapril	Valsartan
Sitagliptina	Linagliptina	Saxagliptina	Labetalol	Enalapril	Fosinopril
Saxagliptina	Alogliptina	Gliburida	Perondopril	Ramipril	Aliskiren
Glimepirida	Glipizida	Dapagliflozina	Trandolapril	Candesartan	Eprosartán
Canagliflozina	Empaglifozina	Exenatida	Irbesartan	Telmisartán	Diltiazem
Dulaglutida	Acarbosa	Miglitol	Felodipino	Metildopa	Espirinolactona
Pioglitoza	Aspart	Lispro	Isradipina	Prazosin	Terazosina
Glulisina	R-cristalina	Glargina	Carvedilol	Clonidina	Hidralazina

Una vez teniendo los principales medicamentos utilizados en el tratamiento de hipertensión y diabetes, las fuentes de información elegidas para la recolección de comentarios fueron las publicaciones de Twitter y comentarios en Facebook, esto debido a que son redes sociales con un gran número de usuarios, así mismo Twitter tiene un API disponible para desarrolladores la cual brinda una gran facilidad para obtener información y da la posibilidad de extraer hasta 2 millones de tweets mensualmente de forma gratuita. Por otro lado, para Facebook es necesario utilizar la librería de Python *facebook_scraper* para obtener los comentarios dadas ciertas URLs de publicaciones. Los términos utilizados para realizar la búsqueda en ambas redes sociales se muestran en la Tabla 8 donde además de buscar los medicamentos mencionados anteriormente, se busca la enfermedad y algunas variantes con palabras extra.

Tabla 8 Conjunto de palabras utilizados para la búsqueda de comentarios en redes sociales.

Palabras de búsqueda
Diabetes, diabetes tipo 2, diabetes tipo 1, diabetes mellitus, medicamentos diabetes, hipertensión, medicamentos hipertensión, metformina, semaglutida, liraglutida, losartan, quinapril, valsartan, sitagliptina, linagliptina, saxagliptina, labetalol, enalapril, fosinopril, saxagliptina, alogliptina, gliburida, perondopril, ramipril, aliskiren, glimepirida, glipizida, dapagliflozina, trandolapril, candesartan, eprosartán, canagliflozina, empaglifozina, exenatida, irbesartan, telmisartán, diltiazem, dulaglutida, acarbosa, miglitol, felodipino, metildopa, espirinolactona, pioglitoza, aspart, lispro, isradipina, prazosin, terazosina, glulisina, r-cristalina, glargina, carvedilol, clonidina, hidralazina.

Para realizar la consulta al API de Twitter es necesario tener un perfil como desarrollador para obtener un token de autenticación y poder consumir el API mediante un *Bearer token* que funciona como autenticación. Esta consulta se realizó creando un programa en Python ya que el API de Twitter solo permite consultar 100

registros por ejecución como máximo. Este programa se encarga de ejecutar la consulta hasta un límite determinado por el usuario, el límite tomado para esta investigación fue de 500 dando como máximo la obtención de 50 mil tweets por ejecución. Así mismo el programa recupera los tweets en formato JSON, con las siguientes claves: *ID*, *text*; donde ID guarda el Id del tweet y *text* guarda el texto contenido en el tweet eliminando saltos de línea que de otra manera romperían la estructura JSON. Adicional después de cada 100 registros se guarda el *next_token* este sirve para determinar si existen más registros y seguir iterando para conseguir los tweets, esto se puede apreciar en la Figura 1.

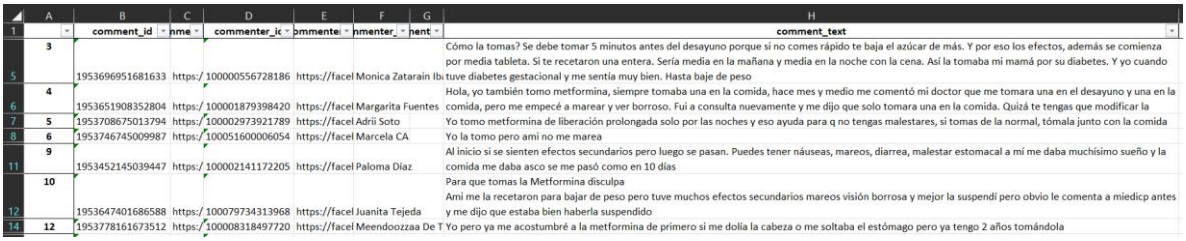

```
{
    "ID": "1625998702388158467",
    "text": "“Metformina”: medicamento para la diabetes que aseguran que retrasa el envejecimiento https://t.co/k0kutqreOb via
    @noticiashouston"
},
{
    "ID": "1625997213578022913",
    "text": "@dominiquemetz al sr que necesita insulina o metformina le regalo. Pero sin cámaras ni notas, ni nada. @todonoticias"
},
{
    "ID": "1625996497501270016",
    "text": "metformina da neo quimica https://t.co/yYCtUEjxv4"
},
{
    "next_token": "b26v89c19zqg8o3fqk718i4599yunc6qd7f29jthfaagt"
},
```


Figura 1 Representación de tweets obtenidos mediante el API de Twitter en formato JSON.

Para el caso de obtención de comentarios de Facebook es necesario proporcionar una serie de URLs en un archivo XLSX tal siguiendo un formato de número y URL tal y como se muestra en la Figura 2, para que con ayuda de la librería *facebook_scraper* se puedan extraer los comentarios proporcionando una gran cantidad de información referente a los mismos, sin embargo para esta investigación lo más relevante es *comment_id* y *comment_text*, el resultado de la obtención de comentarios de Facebook se muestra en la Figura 3.

	A	B	C	D	E	F	G	H	I	J	K
1	id	url									
2	1	https://www.facebook.com/groups/607712920117124/posts/1225481611673582/									
3	2	https://www.facebook.com/groups/1442339342817399/permalink/1953049991746329/									
4	3	https://www.facebook.com/brenda.marin1/posts/pfbid02VEKk663RZy31vtxnrzpfz9zdP1rL77n4qn6rfF8Ui8Ra2zFRGTx9txsqx1VVQynql									
5	4	https://www.facebook.com/groups/197298663940185/permalink/2078508315819201/									
6	5	https://fb.watch/kJcfNyRQvq/									
7	6	https://fb.watch/kJbRLVTA8M/									
8	7	https://www.facebook.com/semaglutida/posts/pfbid0LtCP27qmNpqwVbemXZF3uitaxvwn82EEMFuBBfGyYUDqdpPjWNVKSHbKZtsPPmzol									
9	8	https://www.facebook.com/nutmolecular/posts/pfbid0jzK9ypZ7ZaaXhPGmsqvCrjDQ4MbKZDy2dRpBvFXAFbGmJpqaGVtgebRehT9gcrELl									
10	9	https://www.facebook.com/reel/1196340081026313									
11	10	https://fb.watch/kJffW7UT1K/									
12	11	https://www.facebook.com/farmaciaespecialidaddelsur/posts/pfbid0fk9vaF2CrqQtX3Dhe1J9LNTXovrKBAHqykeThYEsDz5yHjquTcjMYF7qBJ5RL7A1l									
13	12	https://www.facebook.com/DiabetesJuntosxTi/posts/pfbid02C7vzYSmPUUU3dXdTNUy1AGGsLPj8vraobEVPisq6ntAiQ2ckJ8pc45GoJuHiDi3il									
14	13	https://www.facebook.com/farmaciaespecialidaddelsur/posts/pfbid02TXNyGUFcTw5haPHmJrXMQA7HwwxXEnEarroTFbaoi9LumwadmiJPa1eraENTyQZjl									
15	14	https://www.facebook.com/groups/607712920117124/posts/1225481611673582/									

Figura 2 Conjunto de enlaces utilizados como entrada del scraper de Facebook para la obtención de comentarios.



	A	B	C	D	E	F	G	H
1		comment_id	nme	commenter_ic	ommente	nmenter_	hent	comment_text
5	3	1953696951681633	https:/	100000556728186	https://facel	Monica Zatarain Ib		Cómo la tomas? Se debe tomar 5 minutos antes del desayuno porque si no comes rápido te baja el azúcar de más. Y por eso los efectos, además se comienza por media tableta. Si te recetaron una entera. Sería media en la mañana y media en la noche con la cena. Así la tomaba mi mamá por su diabetes. Y yo cuando tuve diabetes gestacional y me sentía muy bien. Hasta baje de peso
6	4	1953651908352804	https:/	100001879398420	https://facel	Margarita Fuentes		Hola, yo también tomo metformina, siempre tomaba una en la comida, hace mes y medio me comentó mi doctor que me tomara una en el desayuno y una en la comida, pero me empecé a marear y ver borroso. Fui a consulta nuevamente y me dijo que solo tomara una en la comida. Quizá te tengas que modificar la
7	5	1953708675013794	https:/	100002973921789	https://facel	Adrii Soto		Yo tomo metformina de liberación prolongada solo por las noches y eso ayuda para q no tengas malestares, si tomas de la normal, tómala junto con la comida
8	6	1953746745009987	https:/	100051600006054	https://facel	Marcela CA		Yo la tomo pero ami no me marea
11	9	1953452145039447	https:/	100002141172205	https://facel	Paloma Díaz		Al inicio si se sienten efectos secundarios pero luego se pasan. Puedes tener náuseas, mareos, diarrea, malestar estomacal a mí me daba muchísimo sueño y la comida me daba asco se me pasó como en 10 días
12	10	1953647401686588	https:/	100079734313968	https://facel	Juanita Tejeda		Para que tomas la Metformina disculpa Ami me la recetaron para bajar de peso pero tuve muchos efectos secundarios mareos visión borrosa y mejor la suspendí pero obvio le comenta a miedicp antes y me dijo que estaba bien haberla suspendido
14	12	1953778161673512	https:/	100008318497720	https://facel	Meendoozzaa De T		Yo pero ya me acostumbré a la metformina de primero si me dolía la cabeza o me soltaba el estómago pero ya tengo 2 años tomándola

Figura 3 Representación de comentarios obtenidos con el scraper de Facebook.

Posteriormente, se realizó la obtención de información de medicamentos esto se hizo mediante un API creada en Laravel que internamente cuenta con un Scraper. Con esta API por medio del nombre del medicamento se puede obtener y registrar la información del medicamento. La información se obtuvo de la página de oficial de la Administración de Alimentos y Medicamentos (FDA, *Food and Drug Administration*) obteniendo las siguientes características:

- Nombre del medicamento
- Como usar el medicamento
- Condiciones o enfermedades de prescripción
- Efectos adversos reportados

Por otra parte, para la obtención de efectos adversos estos se obtuvieron de manera manual de la base de datos SIDER, la cual es una base de datos publica que registra la gran mayoría de los efectos adversos, su definición, así como medicamentos relacionados, sin embargo, solo se encuentra en idioma inglés. Finalmente, la obtención de las alertas sanitarias se realizó mediante una tarea programada creada en Laravel que se ejecuta una vez al día con el fin de obtener las alertas publicadas por la COFEPRIS de manera actualizada. De estas alertas se obtienen las siguientes características:

- Fecha de publicación
- Enlace de la publicación
- Nombre de la alerta sanitaria

3.1.3 Preparación de los datos

Para esta etapa se realizará el formateo de los datos de acuerdo con lo necesario para esta investigación, así como la limpieza de estos para mejorar la calidad y adecuarlos para los siguientes pasos a realizar.

Preprocesamiento de tweets

Este es un paso necesario dada la naturaleza de los comentarios obtenidos del API de Twitter, ya que presentan características como mención de otros usuarios, textos

con emojis, enlaces a otros sitios web, e incluso tweets en otros idiomas. Partiendo de la necesidad de que los tweets deben estar en idioma español y ya que el nombre de las enfermedades y medicamentos es muy parecido en idiomas como el francés, portugués, italiano e incluso inglés, se tuvo que añadir una clave extra a cada objeto JSON generado en la recopilación inicial de datos, esto se logró gracias a la librería de Python *langdetect* la cual fue diseñada por Google y tiene una precisión del 99% para detectar el idioma de un texto, esto se puede apreciar en la Figura 4.


```
{
    "ID": "1624537631857037313",
    "text": "Dios, tuve que suspender la pastilla que me dio la psiquiatra por lo mal que me hace la metformina (hasta acostumbrarme
    a ella y después vuelvo a las dos) y hace 2 días estoy con una irritabilidad que me dan ganas de matarmeeee. Odio estoy de mal
    humor SIN RAZÓN ALGUNAA",
    "language": "es"
},
{
    "ID": "1624528709569970178",
    "text": "#QUIZ - #Diabetes: Quale agente complementare privilegiare con la metformina? #MedTwitter https://t.co/3NCkWfrHKX
    https://t.co/FH73oMAEmm",
    "language": "it"
},
{
    "ID": "1624524052265070593",
    "text": "@ThomssAF Aunque probs una AI si pueda identificar esa 'pastillita roja chiquita' con una buena base de datos.   'Quizá
    querías decir metformina'?",
    "language": "es"
},
{
    "ID": "1624522708540088322",
    "text": "RT @detodonada: Compro un medicamento que se llama Galvus Met Vildagliptina 50 mg / Metformina 850 mg de 56 Comprimidos.
    En las cadenas de…",
    "language": "es"
},
{
    "ID": "1624520538507862016",
    "text": "Não tem remédio e ainda aumentaram o  valor da metformina era   6 reais agora já está 11 reais das que tinha no mercado
    as farmácias estão aproveitando pra lucrar https://t.co/Zych7I9tll",
    "language": "pt"
},
{
    "ID": "1624519744589004807",
    "text": "Sou diabética não tem remédios no farmácia popular certo desde final do ano metformina 850 nao tem está em falta eo povo
    que se vire alguma faz  coisa presidente  #LulaPresidente https://t.co/2suAOJBHPe",
    "language": "pt"
},
```


Figura 4 Conjunto de tweets en formato JSON con clave de idioma añadida mediante librería langdetect.

Después de colocar la clave del idioma a los objetos JSON se procedió a seleccionar únicamente aquellos cuya etiqueta *language* fuera "es" que son los tweets en idioma español, quedando un total de 17035 tweets. Con el fin de realizar una limpieza de los comentarios y mejorar su calidad se siguió lo presentado por Sakhovskiy & Tutubalina (2022), donde se indican los siguientes pasos: los emoticonos o *emojis* deben remplazarse con una palabra que corresponda a lo que describe el emoticono, así mismo las menciones a otros usuarios deben ser remplazados por una mención genérica como @USUARIO y finalmente los enlaces a otros sitios web deben ser remplazados por <LINK> para enmascarar el enlace, todo este proceso se puede visualizar en la Figura 5.

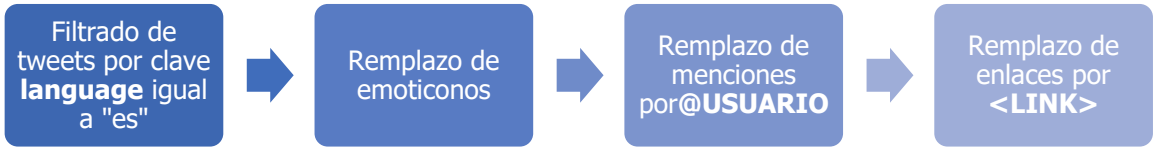


Figura 5 Diagrama de proceso para realizar un preprocesamiento de los datos.

Para concluir con este preprocesamiento se realizó la eliminación de tweets repetidos por medio de una comparación de contenido quedando así un total de 13961 tweets.

Clasificación de comentarios

Clasificación de tweets

La clasificación de tweets se llevó a cabo de manera manual discriminando aquellos tweets que no tenían relación con las palabras buscadas, así como tweets con contenido ofensivo o que abordaran temas no relacionados con la instigación, quedando así un total de 309 tweets, de los cuales 261 no tienen mención de un efecto adverso (no-RAM) y 48 si tienen la mención de un efecto adverso (si-RAM).

Clasificación de comentarios de Facebook

Para la clasificación de comentarios de Facebook se siguieron los mismos criterios que para los tweets, de un total de 576 comentarios obtenidos de Facebook, se obtuvieron un total de 52 comentarios relacionados con los efectos adversos, de los cuales 31 fueron clasificados como no-RAM y 21 fueron clasificados como si-RAM.

Construcción de Corpus

Para la construcción de corpus se integraron los tweets y los comentarios de Facebook en un solo conjunto de datos en formato TSV, con los siguientes campos ID, el texto del comentario y finalmente su clase, así como se muestra en la Figura 6. El corpus final se compone de un total de 361 elementos de los cuales 69 se clasifican como si-RAM y 292 como no-RAM.

```
1627214225310375936 El experto en distribución de medicinas explica que «por ejemplo, dispararon el uso de semaglutida para adelgazar y los diabéticos ahora no la encuentran» <LINK>  0
1627100273079459840 Efecto de la Semaglutida en la Proteína C Reactiva en Adultos con Sobrepeso y Obesidad <LINK> vía @USUARIO  0
1626898923762712576 El Dr. @USUARIO  inicia su intervención señalando que en el análisis post hoc realizado con los estudios SUSTAIN 6 y PIONEER 6 se observa que el beneficio cardiovascular que produce la semaglutida ocurre a lo largo de todos los estadios de función renal #ARC #diabetes <LINK> 0
1626877269456166913 En entorno de vida real en España (estudio REALSEM-SP), la persistencia con semaglutida 1 mg/semana al año de tratamiento fue elevada, con una baja interrupción por efectos adversos, añade la Dra. @USUARIO #ARC en #diabetes 0
1626667633503354881 RT @USUARIO: Semaglutida subcutánea tuvo un efecto beneficioso sobre la esteatosis hepática que fue más allá del control de la glucosa  0
1626666649716772864 Terapias farmacológicas para la obesidad que tenemos que conocer para saber prescribirlas.   #Semaglutida 2.4 mg #Tirzepatide   Todas han probado su eficacia CON estilo de vida adecuado. No hay terapias milagro. <LINK>  0
1626657996783640588 Los pacientes con DM2 y eventos cardiovasculares que usan semaglutida tienen una reducción del 55% de riesgo de sufrir un evento cerebrovascular, afirma el Dr. @USUARIO #ARC en #diabetes  0
1626605538053062658 Aunque la búsqueda de terapias para el NASH debe seguir en marcha, explica la Dra. @USUARIO, debemos aumentar el uso de las terapias que han demostrado beneficio (liraglutida, semaglutida, pioglitazona) #ARC en #diabetes    0
1626585841031327747 RT @USUARIO: Semaglutida subcutánea tuvo un efecto beneficioso sobre la esteatosis hepática que fue más allá del control de la glucosa  0
```

Figura 6 Fragmento de corpus en formato TSV.

Posteriormente para preparar la entrada de datos a los modelos de redes neuronales y de aprendizaje máquina se procedió darle al corpus los formatos necesarios. Para las redes neuronales se necesita identificar el objetivo de la oración esto se realizó mediante un programa el Python el cual remplaza el objetivo de la oración, que en este caso es el nombre del medicamento, por el token *T*, y finalmente se segmenta en filas, donde la primera fila es el contenido con el token, la segunda el nombre del medicamento y la tercera su clase esto se puede apreciar en la Figura 7.

```
@USUARIO La $T$ no hace milagros, no hay que abusar de esas personas mayores, que irresponsabilidad de parte de ustedes.
losartan
0
@USUARIO @USUARIO Mi padre le dio un infarto el 28 de diciembre tomaba esos medicamentos hace más de un año por más que le pelee eso pero obvio dice es que no tengo para
comprar en otro sitio. Toma actualmente $T$ potásico de 100mg dos veces al día, carvedilol de 12.5mg, nifedipino LP, clopid
losartan
0
amor/odio a la $T$
metformina
0
El $T$ es el primer anticuerpo que retrasa la aparición de la diabetes tipo 1 en al menos 2 años en pacientes de alto riesgo. El primero por lo tanto que retrasa la aparición
de una enfermedad autoinmune. Pero es terriblemente caro. <LINK> <LINK>
Teplizumab
0
Eficacia de la $T$ entre las medicaciones basales de la diabetes: análisis preespecificado a partir del estudio DAPA-CKD <LINK> PMID 36662635
dapagliflozina
0
RT @USUARIO: (1/2) SÍ, el $T$ puede subir la tensión y debe ser consumido con precaución o en dado caso evitado, siempre bajo…
Ibuprofeno
1
```

Figura 7 Fragmento de corpus el cual ha sido adecuado al formato de entrada de las redes neuronales RAM e IAN.

De acuerdo con la Figura 8 y para el algoritmo de aprendizaje máquina requiere un formato JSON de entrada con las claves *text* que es el contenido del comentario, *label* que es la clase, *entity* que es el nombre del medicamento y finalmente el *id* que es el identificador del comentario.

```
{"text":"“Metformina”: medicamento para la diabetes que aseguran que retrasa el envejecimiento <LINK> via @USUARIO", "label":"Unknown",
"entity":"“Metformina”", "id":19777804}
{"text":"Me a tomar una metformina porque todo esto me hace subir el azúcar", "label":"Unknown", "entity":"metformina", "id":15217097}
{"text":" Uso de benzodiazepinas pastilla  e hipertensión en urgencias  alarma .    Revisión publicada en @USUARIO @USUARIO resumido en
@USUARIO.   ¿Hay beneficio clínico estetoscopio ? Pueden utilizarse como un antihipertensivo en algún caso? #CardioTwitter #MedTwitter
<LINK>", "label":"Unknown", "entity":"benzodiazepinas", "id":18893818}
{"text":"@USUARIO Un familiar muy cercano..tuvo q cambiar el losartan de laboratorio reconocido x ese medicamento hindú.xq no tenía
dinero ..le causo un infarto y fallecio...Así q Ud tiene razón y conozco a otras personas con igual resultado   ..esos medicamentos indu
estan matando", "label":"Adverse", "entity":"losartan", "id":15921607}
```

Figura 8 Fragmento de corpus el cual ha sido adecuado al formato de entrada del algoritmo SVM.

3.1.4 Modelado

De acuerdo con la metodología CRISP-DM en este apartado se describirán las técnicas utilizadas para cumplir con los objetivos definidos en el Entendimiento del Negocio. Así mismo se hará la descripción de las pruebas con las técnicas elegidas y la generación del modelo.

Técnicas de modelado

Dada la complejidad del problema que la identificación de presencia o ausencia de efectos adversos el utilizar algoritmos de aprendizaje máquina y redes neuronales es de gran relevancia, así mismo de acuerdo con la revisión del estado del arte se

describen a continuación las técnicas elegidas que han dado buenos resultados en este problema en específico.

SVM

El algoritmo de Máquinas de Vectores de Soporte es un método de tipo de aprendizaje automático el cual realiza un aprendizaje supervisado. Este método es utilizado principalmente para la clasificación o regresión de grupos de datos. SVM funciona mediante la creación de un hiperplano en un espacio multidimensional para dividir las diferentes clases. Para realizar la clasificación binaria SVM trabaja de manera iterativa por lo cual repite el proceso de creación de hiperplano tratando de buscar el margen máximo entre ambas clases para obtener la mejor clasificación del conjunto de datos maximizando la distancia entre los puntos de datos de ambas clases, este proceso se puede apreciar en la Figura 9. Mientras que, para realizar la regresión, SVM busca la función que mejor aproxima los puntos de datos, minimizando la suma de las distancias entre los puntos y la función. Así mismo, SVM puede manejar tanto múltiples variables continuas como categóricas.

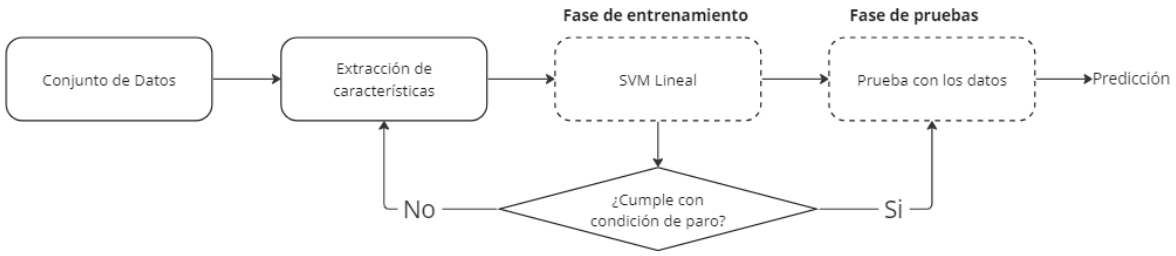


Figura 9 Diagrama que muestra la ejecución del algoritmo SVM.

RAM

Esta es una red neuronal de atención recurrente propuesta por Peng et al. (2017) para el análisis de sentimientos que consta de cinco módulos los cuales son entrada de datos, módulo de memoria, memoria ponderada por posición, módulo de atención recurrente y salida, tal y como se evidencia en la Ilustración 10. El módulo de entrada recupera los vectores de palabras de una tabla embebida generada por un método no supervisado como GloVe o CBOW. El módulo de memoria utiliza la red neuronal *Deep Bidirectional LTSM* para construir la memoria que registra toda la información que se leerá en módulos posteriores. El módulo de memoria ponderada por posición se utiliza para producir una memoria de entrada para cada objetivo asignando un mayor peso a las palabras más cercanas al objetivo. El módulo de atención recurrente se utiliza para actualizar la memoria con una suma ponderada de la memoria y un vector de consulta generado a partir de la salida anterior con el

fin de construir de manera adecuada la clasificación de sentimientos. Por último, el módulo de salida utiliza una función *softmax* para predecir la polaridad del sentimiento de la palabra objetivo.

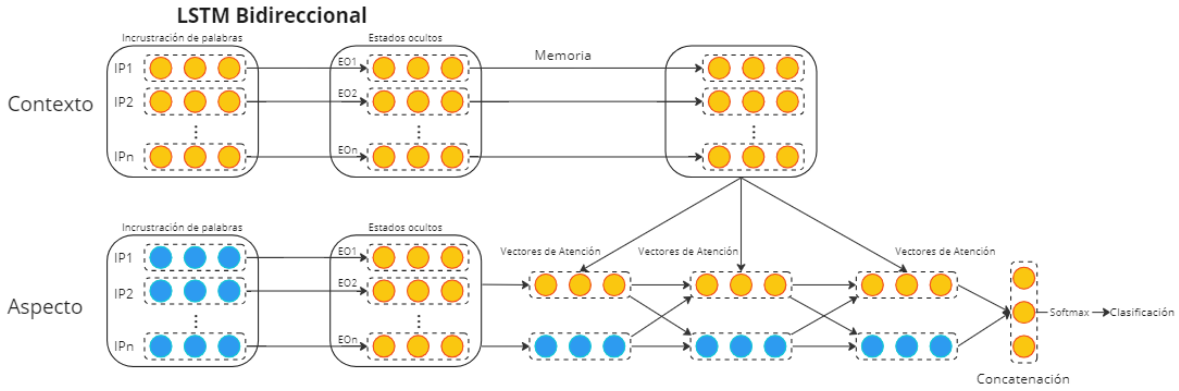


Figura 10 Diagrama de arquitectura de red neuronal RAM.

IAN

Este modelo fue propuesto por (Ma et al. (2017) es una red neuronal de atención interactiva para la clasificación de sentimientos con un enfoque en aspectos. Este modelo se compone de dos partes, utiliza redes LTSM para modelar el objetivo y el contexto de manera interactiva obteniendo estados ocultos de las palabras. El mecanismo de atención se utiliza para obtener la información importante tanto del contexto como del objetivo. Así mismo, después de calcular los pesos del contexto y del objetivo estos vectores se concatenan y alimentan a una función *softmax* para la clasificación del sentimiento a nivel de aspecto, este proceso se representa en la Figura 11.

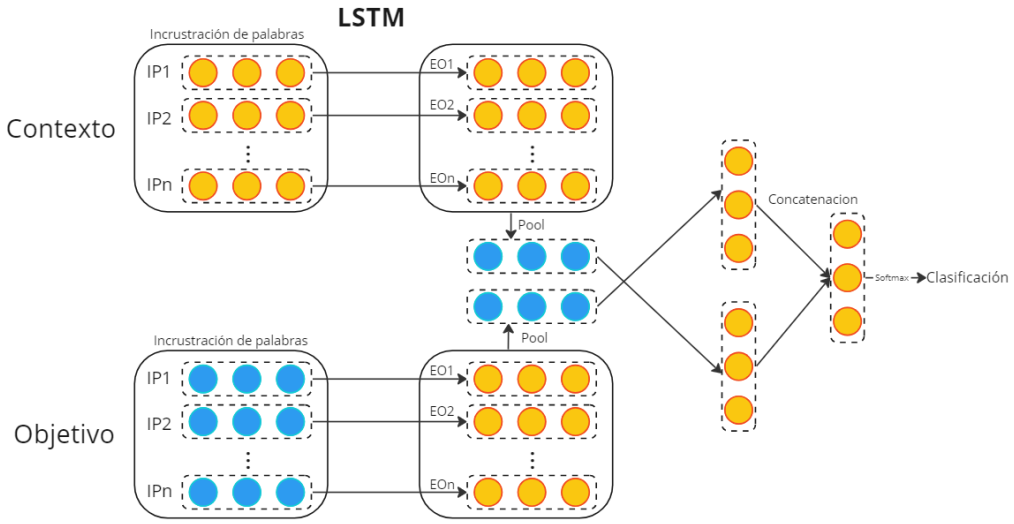


Figura 11 Representación de arquitectura de la red neuronal IAN.

Prueba de técnicas

Para poder realizar las pruebas con las técnicas elegidas es necesario dividir los datos en datos de entrenamiento y datos de prueba. Esto se realizó respetando una proporción 70/30 tal y como se aprecia en la Tabla 9. Los datos fueron elegidos de manera aleatoria conservando la proporción de aproximadamente 19% de efectos adversos presentes en cada conjunto de datos.

Tabla 9 Resumen de estadísticas del conjunto de datos.

	Entrenamiento	Prueba	Todos
Comentarios	252	109	361
Comentarios con RAM	48	21	69
Comentarios sin RAM	204	88	292
Proporción si-RAM	19.04%	19.26%	19.11%

Para realizar las pruebas de las técnicas elegidas se inició con SVM se utilizó el algoritmo desarrollado por I. Alimova & Tutubalina (2018) el cual está enfocado en la clasificación de efectos adversos. Como parámetros de entrada se utilizaron dos vectores de emoticones con representación en caracteres ASCII con polaridad positiva y negativa. Así mismo se ingresó un vector de medicamentos en español presentados en la Tabla 7. Para el vector de palabras incrustadas se utilizó el corpus *Spanish Billion Word Corpus* creado por Cardellino, (2016) y entrenado con Word2Vec el cual tiene 1.4 mil millones de palabras etiquetas en español y su dimensión de vectores de 300. Finalmente se ingresó un diccionario de efectos adversos obtenidos de SIDER el cual cuenta con el registro de más de 13 mil RAM y el cual fue traducido a idioma español mediante un algoritmo de Python y se utilizó la librería *translate* para lograr este fin.

Para las redes neuronales IAN y RAM se utilizaron los parámetros probados por (I. S. Alimova & Tutubalina, 2020) ya que de acuerdo con los experimentos realizados son los que producían los mejores resultados. Se consideran los siguientes: el paso de como máximo 15 épocas para el aprendizaje, el número de estados ocultos es de 300, así mismo el ratio de aprendizaje se estableció en 0.01 y el parámetro de regularización l2 se estableció en 0.001. Estos modelos requieren de vectores de palabras embebidos para obtener los valores de estas y una clasificación más exacta por lo cual se decidió utilizar el corpus *Spanish Unannotated Corpora* entrenado con

el algoritmo *FastText* creado por (Cañete, 2019) que tiene clasificadas y etiquetadas más de 2.6 mil millones de palabras en español y la dimensión de sus vectores es de 300.

3.1.5 Evaluación

En este apartado se describirán los resultados obtenidos de la aplicación de las técnicas para crear el modelo elegido, así mismo se describirá si se alcanzaron los objetivos plateados de acuerdo con el entendimiento del negocio.

Descripción de resultados del modelado

Las tres técnicas elegidas para generar el modelo fueron evaluadas de acuerdo con los siguientes parámetros a la precisión (P) la cual tiene la se muestra en la Formula 1 y se encarga de medir que tan preciso es el ratio de elementos relevantes verdaderos entre los elementos reales recuperados.

$$(1)\quad P = \frac{Elementos\ verdaderos\ positivos}{Elementos\ Verdaderos\ positivos + Verdaderos\ Negativos} = \frac{VP}{(VP + VN)}$$

Así mismo se midió las exhaustividad o sensibilidad (R) la cual mide la fracción de elementos relevantes verdaderos entre elementos relevantes, así como se aprecia en la Formula 2. Además, con las medidas anteriores es posible calcular el Valor-F que es una medida para probar que tan preciso es el modelo y que no debe confundirse con P, esto se puede apreciar en la Formula 3. Finalmente, en la Formula 4 se muestra otra medida a tomar en cuenta la cual es la exactitud (Acc) que es una medida utilizada para medir que tan bien se identifica o excluyen los elementos en una clasificación binaria.

$$(2)\quad R = \frac{Elementos\ verdaderos\ positivos}{Elementos\ Verdaderos\ positivos + Falsos\ Positivos} = \frac{VP}{(VP + FP)}$$

$$(3)\quad F1 = 2 * \frac{Precisión * Exhaustividad}{Precisión + Exhaustividad} = 2 * \frac{P * R}{(P + R)}$$

$$(4)\quad Acc = \frac{Verdaderos\ positivos + verdaderos\ negativos}{Verdaderos\ positivos + Verdaderos\ Negativos + Falsos\ Positivos + Falsos\ Negativos} = \frac{VP + VN}{(VP + VN + FP + FN)}$$

En la Tabla 10 se pueden apreciar los resultados de las técnicas aplicadas. De acuerdo con los resultados obtenidos se puede notar que la red neuronal RAM es la que da mejores resultados correspondiendo con lo establecido en el estado del arte

en (I. S. Alimova & Tutubalina, 2020) al obtener un Valor-F de 79% y una exactitud del 86% en la clasificación del corpus.

Tabla 10 Resultados obtenidos al aplicar SVM, RAM e IAN sobre el corpus generado.

	P	R	F1	Acc
SVM	0.76	0.79	0.77	0.85
RAM	**0.77**	**0.82**	**0.79**	**0.86**
IAN	0.74	0.79	0.76	0.84

Evaluación de objetivos

De acuerdo con los objetivos planteados en el apartado de entendimiento del negocio, se puede afirmar el cumplimiento de todos los objetivos, así como se describe a continuación:

- Se logró construir un corpus de comentarios que incluyen la mención de efectos adversos.
- Se recolectó información de medicamentos, efectos adversos y sus características.
- En el apartado de entendimiento de datos, se obtuvieron las alertas sanitarias de la COFEPRIS por medio de una tarea programada.
- Finalmente se generó un modelo que es capaz de identificar en comentarios la presencia o ausencia de efectos adversos, tomando en cuenta los resultados del modelado se eligió la red neuronal RAM debido a que fue la que mostró mejores resultados.

3.1.6 Despliegue

Para el despliegue de lo desarrollado con la metodología CRISP-DM el modelo desarrollado se integrará a la parte lógica del sistema web y aplicación móvil. La descripción a detalle sobre la integración será cubierta en el apartado de la metodología de desarrollo.

3.2 Metodología de desarrollo

Como se mencionó en el Capítulo 2, para el desarrollo de la metodología de desarrollo se seguirá la metodología SCRUM dadas las características de esta. A continuación, se describirá el desarrollo de la metodología y como se aplica al desarrollo del proyecto.

3.2.1 Historias de Usuario

De acuerdo con la metodología SCRUM en el análisis de requisitos es necesario definir las historias de usuario, respondiendo a las siguientes preguntas ¿cómo?,

¿quién?, ¿para?, así como la definición de criterios de aceptación y su puntuación estimada.

Tabla 11 Descripción de HU para creación base del Proyecto.

	Historia de Usuario 1: Creación estructura base de proyecto
Como	Desarrollador
Quiero	Crear una aplicación web y una aplicación móvil
Para	Que el usuario pueda consultar información de efectos adversos producidos por medicamentos, sus características y conocer las ultimas alertas sanitarias emitidas por la COFEPRIS.
Criterios de Aceptación	C1. Creación de proyecto Laravel para *backend*. C2. Proyecto en React para aplicación Web. C3. Proyecto en React Native para aplicación Móvil.
Puntos estimados	3

Tabla 12 Descripción de HU para extracción de comentarios relacionados a efectos adversos.

	Historia de Usuario 2: Extracción y guardado de comentarios con menciones de RAM
Como	Desarrollador
Quiero	Un método de extracción de comentarios
Para	Construir un corpus y modelo identificador de RAM
Criterios de Aceptación	C1. Crear módulo encargado de la extracción de datos del API Twitter y comentarios de Facebook. C2. Creación de BD y definición de estructura para guardar JSON de tweets y comentarios de Facebook. C3. Crear un módulo para el preprocesamiento de los tweets y comentarios de Facebook.

	C4. Crear modelo identificador de RAM. C5. Agregar información de los nuevos comentarios a las características de los medicamentos.
Puntos estimados	8

Tabla 13 Descripción de HU para extracción de información complementaria para el sistema.

	Historia de Usuario 3: Extracción y guardado de datos de medicamentos y alertas sanitarias COFEPRIS
Como	Desarrollador
Quiero	Una forma de consultar medicamentos y conocer las últimas alertas sanitarias de la COFEPRIS
Para	Poder contar con la información relevante para el desarrollo del sistema web y aplicación móvil
Criterios de Aceptación	C1. Crear módulo encargado de la extracción de datos de medicamentos. C2. Crear módulo encargado de extraer alertas sanitarias y guardar información. C3. Crear APIs de consulta a esta información.
Puntos estimados	5

Tabla 14 Descripción de HU para creación de aplicación web.

	Historia de Usuario 4: Creación de vistas en aplicación web
Como	Usuario
Quiero	Poder acceder mediante la web a la información relacionada con los efectos adversos causados por medicamentos

Para	Conocer la información de los medicamentos, sus efectos adversos y conocer las últimas alertas sanitarias
Criterios de Aceptación	C1. Vista de listado, búsqueda y detalle de medicamentos. C2. Vista de listado, búsqueda y detalle de efecto adverso. C3. Vista para mostrar alertas sanitarias. C4. Vista para mostrar detalle de alertas sanitarias en PDF.
Puntos estimados	5

Tabla 15 Descripción de HU para creación de aplicación móvil.

	Historia de Usuario 5: Creación de vistas en aplicación móvil
Como	Usuario
Quiero	Poder acceder mediante una aplicación móvil a la información relacionada con los efectos adversos causados por medicamentos
Para	Conocer la información de los medicamentos, sus efectos adversos y conocer las últimas alertas sanitarias
Criterios de Aceptación	C1. Vista de listado, búsqueda y detalle de medicamentos. C2. Vista para mostrar alertas sanitarias. C3. Vista para mostrar detalle de alertas sanitarias en PDF.
Puntos estimados	5

3.2.2 Descripción de APIs y tarea programada

En este apartado se describirán las APIs desarrolladas las cuales se encargan de manipular los datos en la base de datos no relacional y la tarea programada para la obtención de alertas sanitarias, así como la descripción de su uso y documentación para esto se utilizó el documentador *Swagger*, el cual es una potente herramienta

para generar documentación de APIs, mostrando los puntos de acceso de las mismas, los parámetros requeridos y las posibles respuestas obtenibles.

El desarrollo de las APIs de este proyecto fue de vital importancia ya que es la lógica e información para el funcionamiento del sistema web y la aplicación móvil. Dentro de las APIs desarrolladas se encuentran 3 grandes grupos, los cuales son Alertas Sanitarias, Efectos Adversos y Medicamentos, tal y como se muestra en la Figura 12. Estos grupos cuentan con sus propias funcionalidades y particularidades.

Figura 12 Documentación desarrollada con Swagger y agrupación de APIs por tipo de categoría.

Alertas Sanitarias

Este grupo de APIs se centra en obtener la información de las alertas sanitarias (Figura 13), la primer API /api/alertas-sanitarias la información de las alertas sanitarias delimitadas por los parámetros de limite, página, ordenamiento por distintos campos, además de presentar un orden ascendente o descendente. Para la segunda API /api/pdf-alerta-sanitaria esta retorna el PDF en el formato denominado base 64, esto debido a que es la manera más sencilla de consultar y retornar archivos en respuestas JSON. Finalmente, la tercer API /api/alerta-sanitaria/{id} es la consulta de una alerta sanitaria dado su ID. Tal y como se puede apreciar en la Figura 14 se muestran los parámetros opcionales para el consumo del primer API, así como respuestas de ejemplo que brindan información de la respuesta esperada y como es igual al resultado obtenido, esto es común para el resto de APIs de los demás grupos variando únicamente en los campos mostrados dada la información de cada documento de la base de datos.

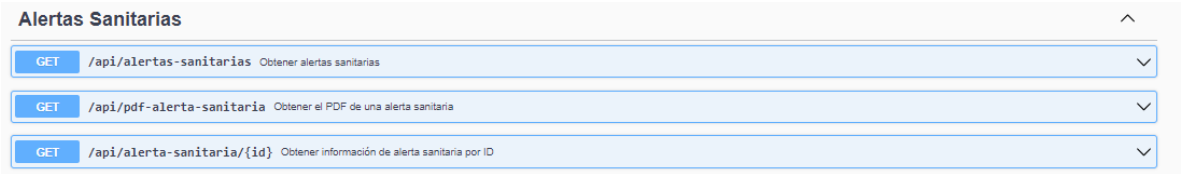


Figura 13 Grupo de APIs que implementan funcionalidades respecto a las alertas sanitarias.

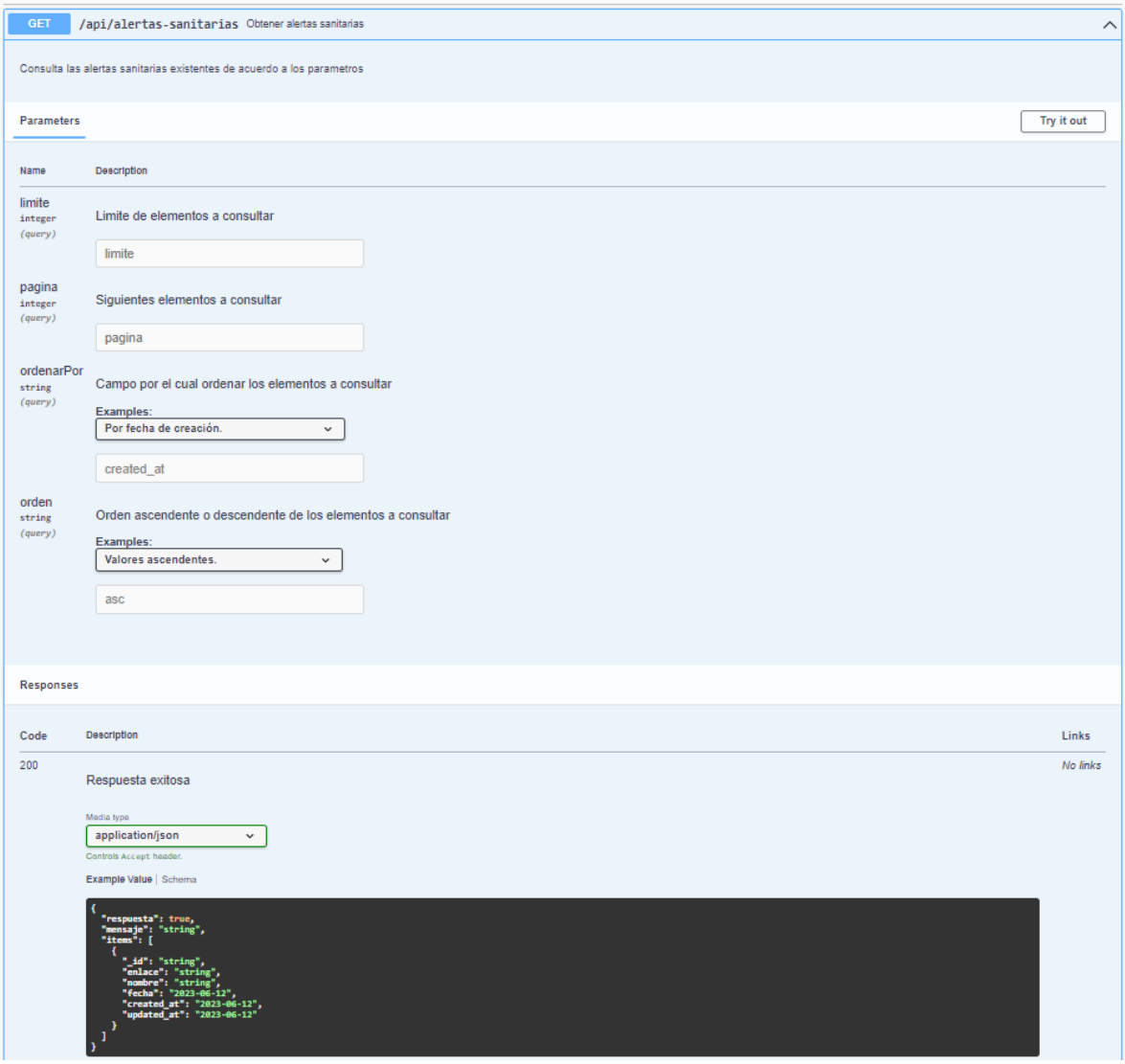


Figura 14 Representación a detalle de documentación de Swagger para API: muestra de punto de acceso, muestra de parámetros y respuestas obtenibles.

Efectos Adversos

Para este grupo de APIs manejan la información de los efectos adversos, la primer API /api/efectosAdversos permite obtener la información completa de los efectos adversos de acuerdo con los parámetros establecidos. La segunda API /api/registrarEfectosAdversos de tipo POST permite registrar nuevos efectos

adversos de acuerdo con su nombre y descripción. Por último, la tercer API /api/efectoAdverso/{id} consulta la información de un efecto adverso en específico de acuerdo con su identificador, los puntos de acceso se pueden observar en la Figura 15.

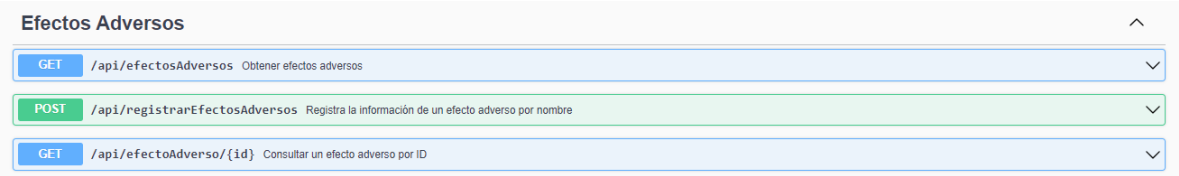


Figura 15 Grupo de APIs que implementan funcionalidades respecto a los efectos adversos.

Medicamentos

Tal y como se aprecia en la Figura 16 este grupo cuenta con cinco APIs distintas, las APIs /api/medicamentos, /api/consultarMedicamento y /api/medicamento/{id}, son parecidas a las descritas en los grupos anteriores, sin embargo, para las APIs /api/consultarMedicamentoURL y /api/registrarMedicamento, estas registran el medicamento dado un enlace o el nombre de un medicamento, respectivamente. Esto lo realizan mediante un scraper hacia la página oficial de la FDA en español y procesan la información para darle el formato correcto y guardarla en la base de datos del proyecto, el proceso llevado a cabo por estas APIs se visualiza en el diagrama de la Figura 17.

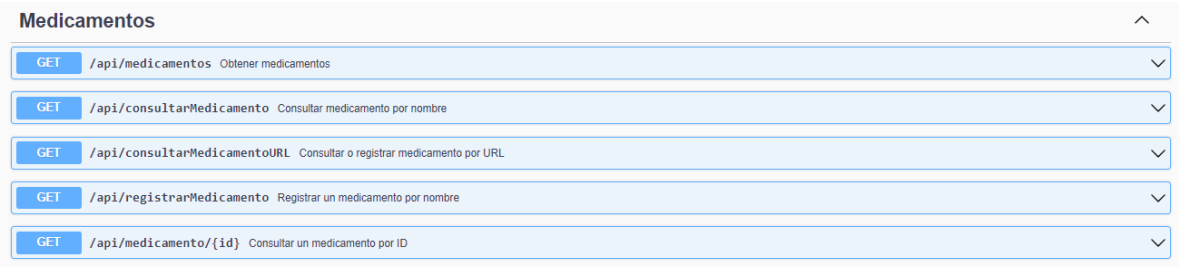


Figura 16 Grupo de APIs que implementan funcionalidades respecto a los medicamentos.

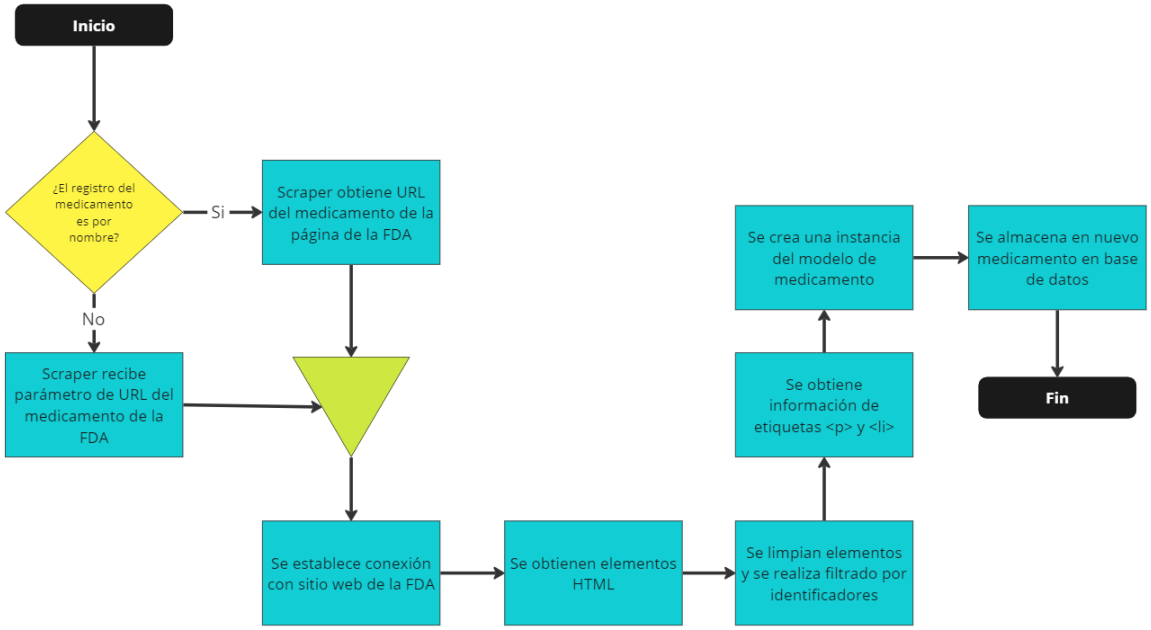


Figura 17 Diagrama de proceso para extraer y guardar información de nuevos medicamentos.

Tarea programada

Esta tarea programada se ejecuta una vez al día a la media noche con el fin de recopilar la información de las alertas emitidas por la COFEPRIS y aunque puede configurarse para ejecutarse más veces al día y a distintas horas esto no es necesario debido a que la COFEPRIS no publica de manera diaria nuevas alertas sanitarias, sino que estas se publican cada que algún caso sanitario referente a medicamentos o productos es reportado. Ya que la COFEPRIS no tiene ningún API o forma de consultar la información de las alertas sanitarias, esta tarea programa cuenta internamente con un scraper que se encarga de recopilar y procesar la información de las alertas sanitarias para almacenar la información en la base de datos del proyecto para posteriormente poder mostrársela al usuario, este proceso se puede evidenciar en la Figura 18.

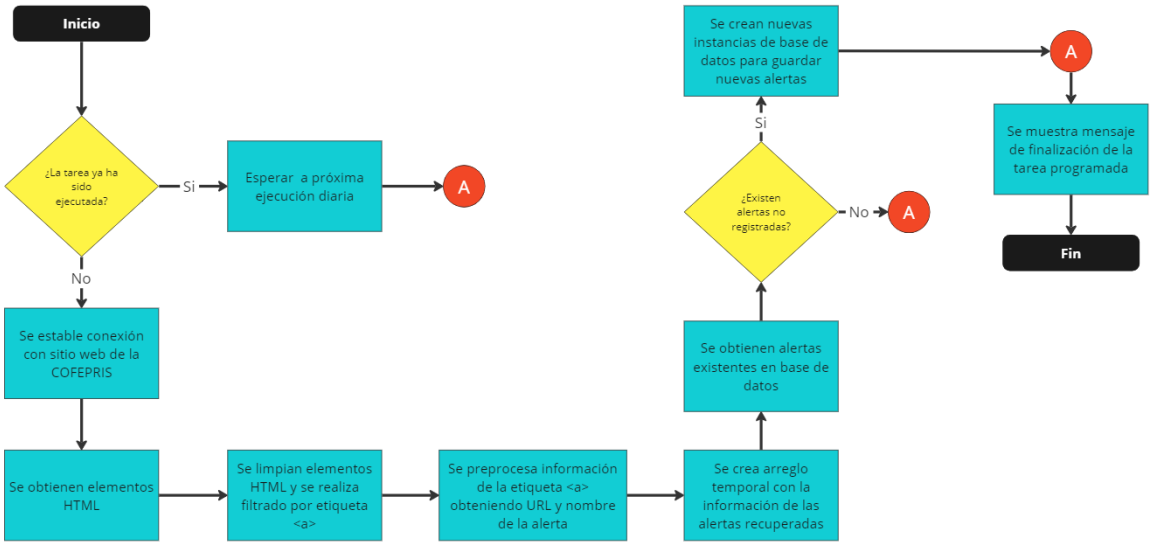


Figura 18 Diagrama de proceso que ejecuta la tarea programada para extraer y registrar nuevas alertas sanitarias.

3.2.3 Diseño de mockups

Continuando con la metodología de desarrollo, para la creación de cualquier sistema es de vital importancia modelar el diseño del mismo, tanto para la aplicación web como la aplicación móvil con el fin de visualizar el diseño del sistema de manera general y tener un primer acercamiento al mismo, así como tener en cuenta las necesidades del usuario y asegurar el que se brinde una buena experiencia de usuario con el uso del sistema. Los mockups fueron diseñados con la herramienta miro en su versión gratuita la cual brinda distintas herramientas para la creación de los mismos. Como se muestra en la Figura 19, esta es la página principal de la aplicación web en la cual se presentan los medicamentos, efectos adversos y alertas sanitarias, así como una barra de navegación la cual estará presente a lo largo de toda la aplicación.

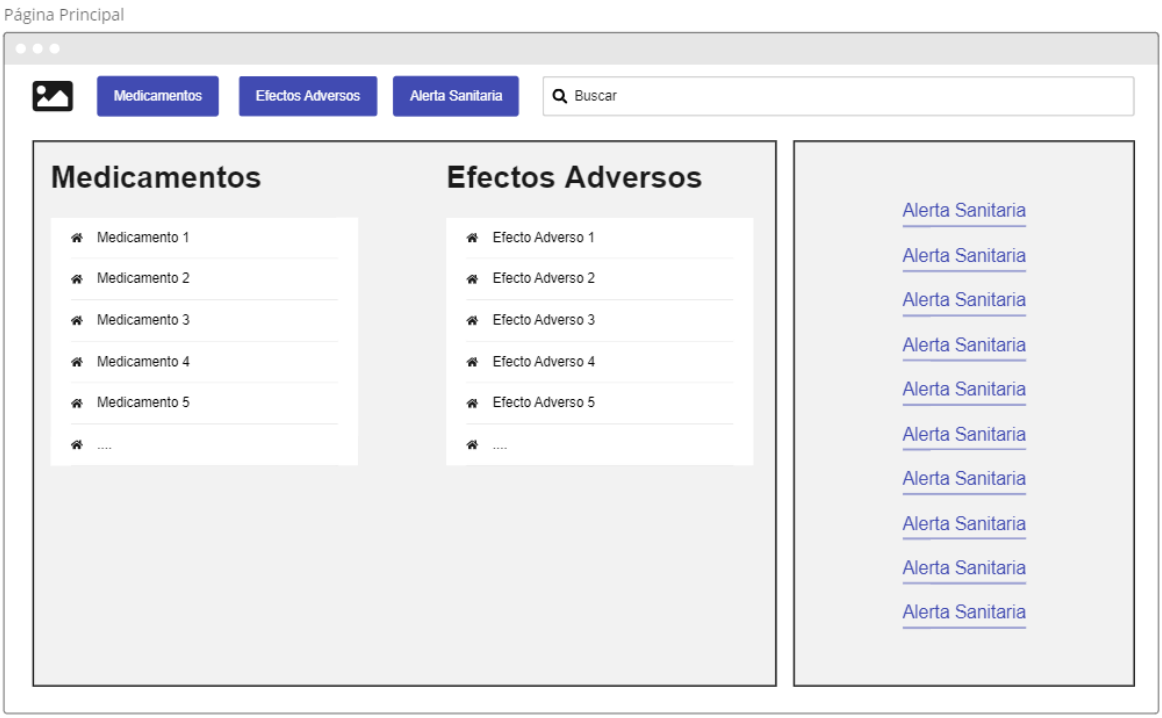


Figura 19 Maquetación de página principal del sistema Sys-RAM.

Pasando con el módulo de los medicamentos se muestra la lista agrupados por su inicial y ordenados alfabéticamente, además de mostrar brevemente la forma de uso de cada medicamento del grupo seleccionado como una vista previa antes de mostrar el detalle del medicamento. Para el detalle del medicamento se muestran principalmente sus efectos adversos oficialmente reportados, así como aquellos descubiertos que no han sido reportados oficialmente, así como mostrar el detalle completo de su condición de uso y forma de uso, esto se puede apreciar en las Figura 20 e Figura 21.

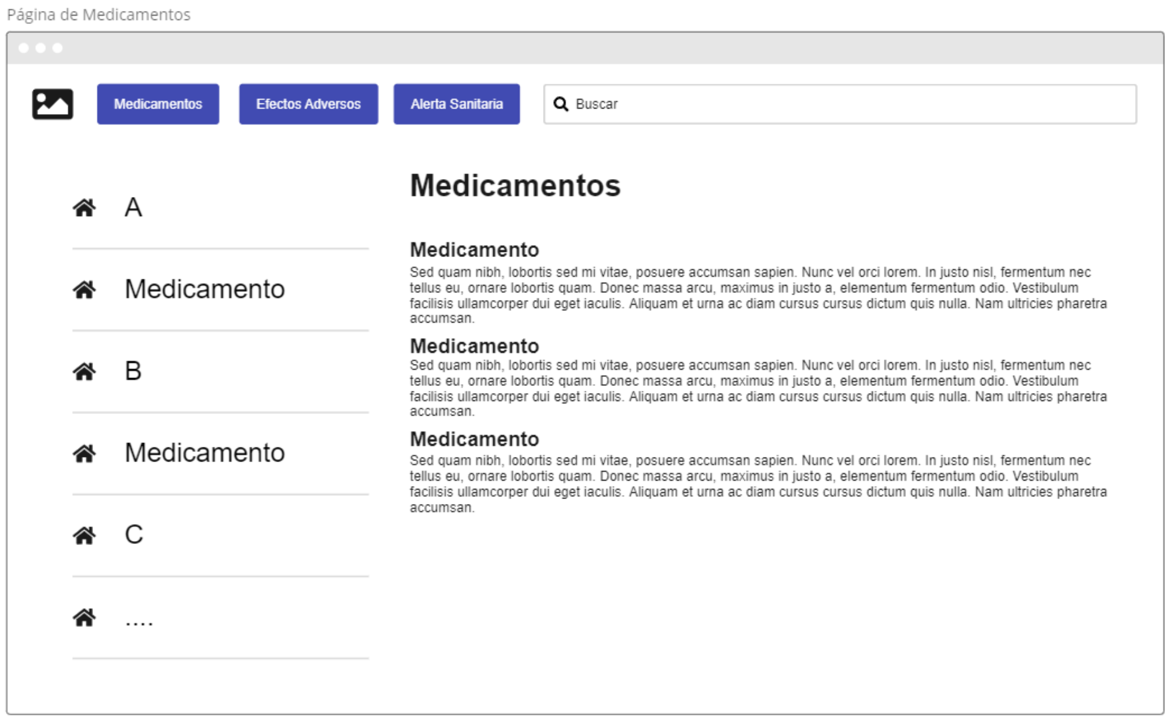


Figura 20 Maquetación de página principal de medicamentos.

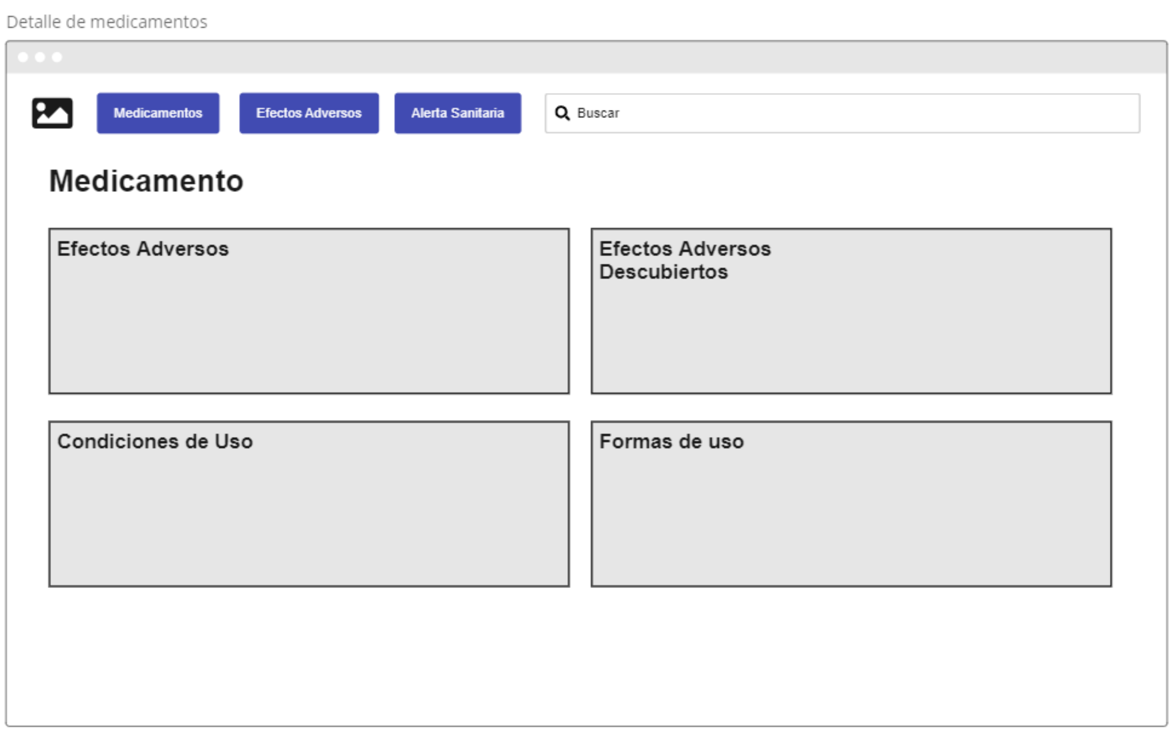


Figura 21 Maquetación de página de detalle de medicamentos.

Así mismo los mockups de la aplicación móvil muestran en primera instancia la lista de medicamentos, desde la cual se puede acceder a mostrar el detalle del medicamento, tal y como se observa en la Figura 22.

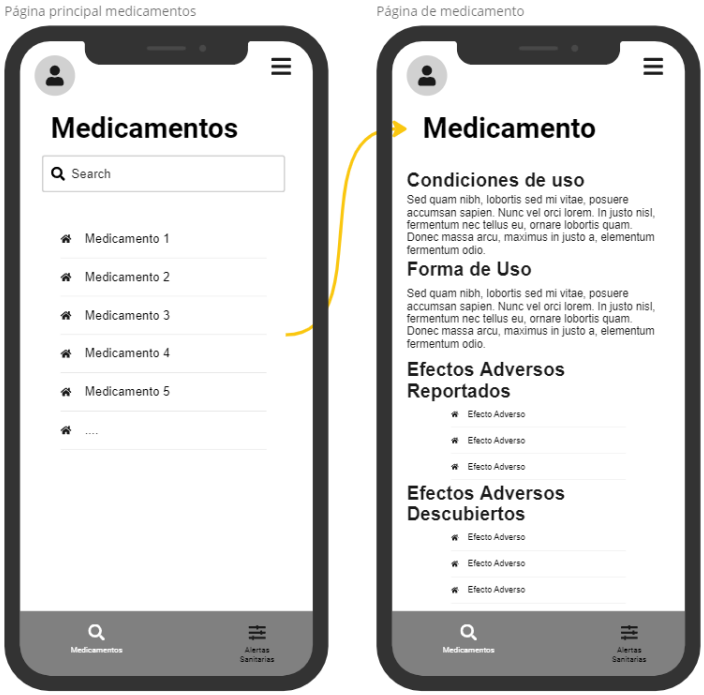


Figura 22 Maquetación para aplicación móvil tanto del listado de medicamentos como el detalle del medicamento.

En la Figura 23 se muestra la página de Efectos Adversos, está página es similar a la de medicamentos se agrupan los efectos adversos por su letra inicial y se muestra una vista previa de la descripción del efecto adverso. Así mismo en la Figura 24 se muestra el detalle del efecto adverso, así como su descripción completa. Este apartado no se modela para la aplicación móvil ya que se busca que esta aplicación cuente solo con la información más importante centrándose en los medicamentos y alertas sanitarias.

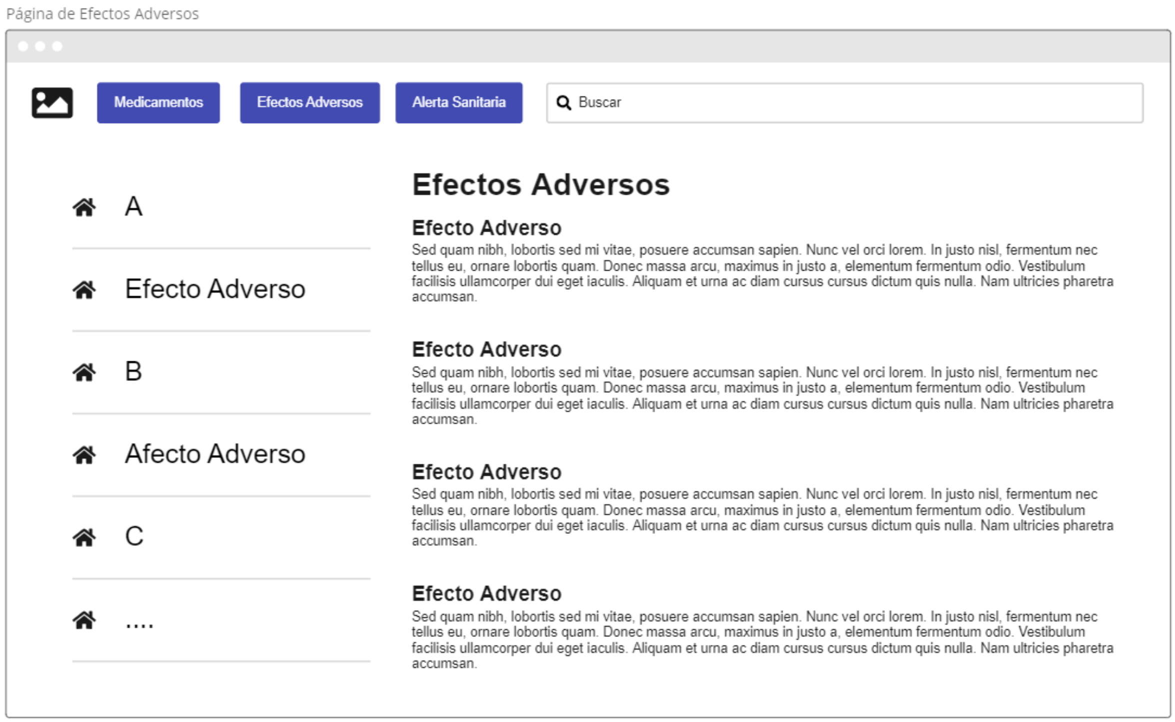


Figura 23 Maquetación para mostrar página que lista los efectos adversos.

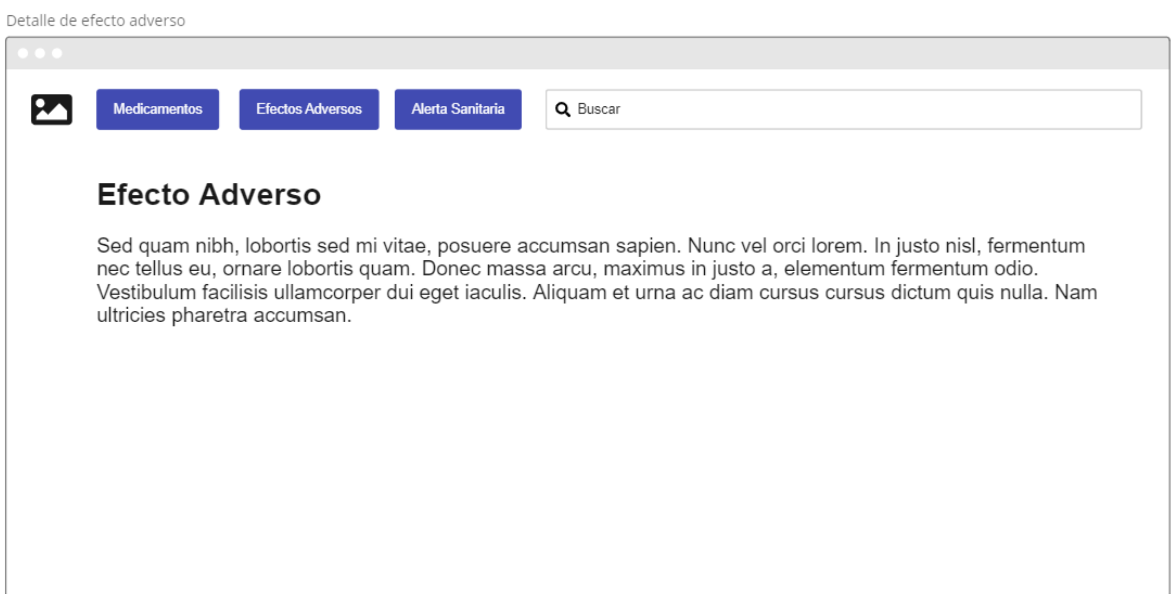


Figura 24 Maquetación para página de detalle de efecto adverso.

Finalmente, para el mockup de las alertas sanitarias se propone una vista donde se listen las alertas sanitarias existentes y se muestre su PDF asociado para que el usuario pueda revisar a detalle el mismo, así como se muestra en la Figura 25. Por otro lado, para la aplicación móvil se modela como el módulo de medicamentos,

mostrando un listado de las alertas sanitarias y posteriormente mostrando el detalle de la misma tal y como se muestra en la Figura 26.

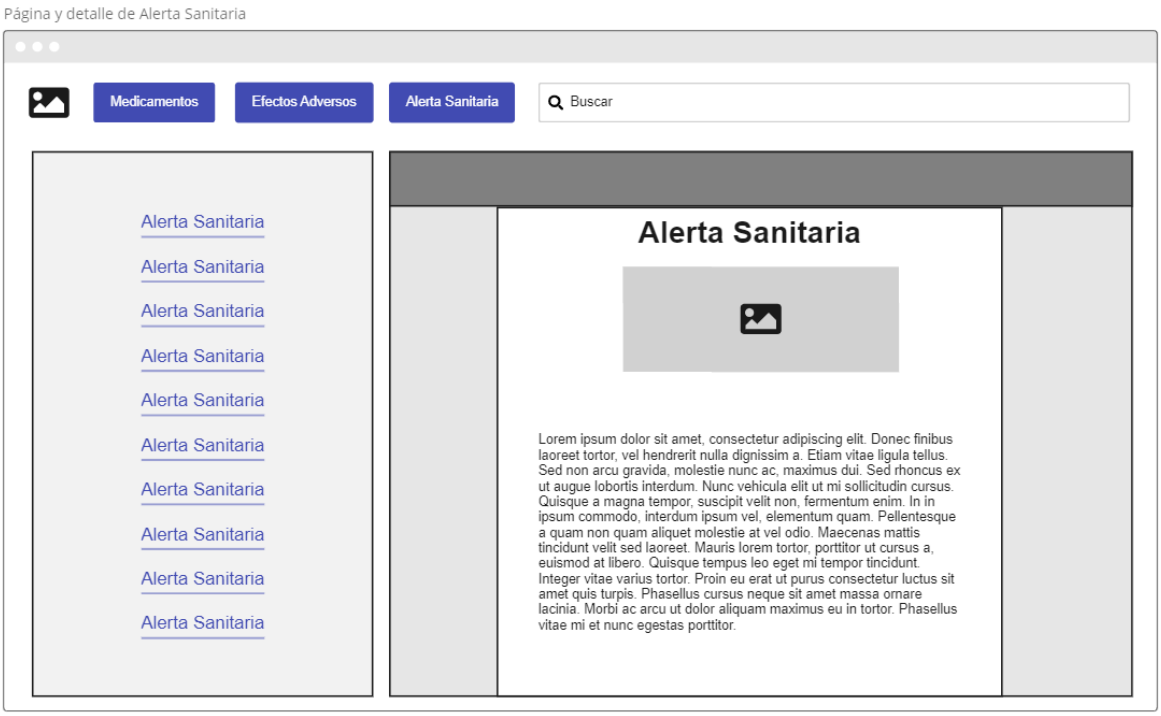


Figura 25 Maquetación de página para visualizar alertas sanitarias.

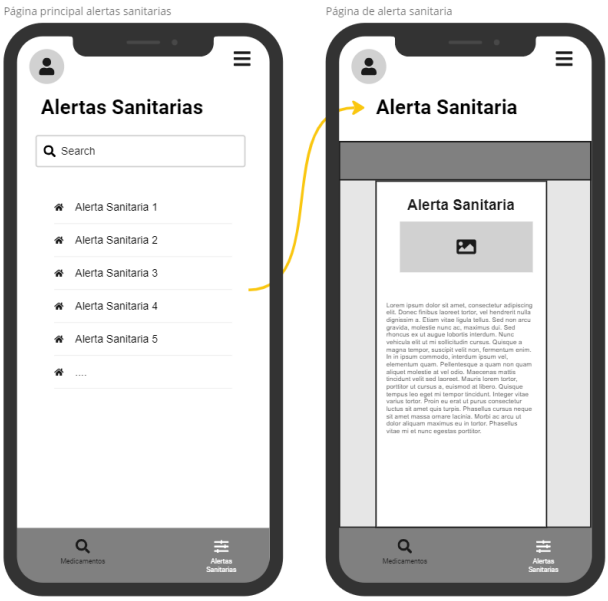


Figura 26 Maquetación de aplicación móvil para visualizar listado de alertas sanitarias y detalle de la misma.

3.2.4 Vistas del sistema

A continuación, se mostrarán las vistas del sistema desarrollado de acuerdo con los diseños creados de los mockups, así como una explicación más detallada del funcionamiento de cada sección.

Vistas de aplicación web

En la Figura 27 se muestra la página principal de la aplicación web con los apartados de medicamentos, efectos adversos y alertas sanitarias. Para el apartado de medicamentos se muestran los últimos 100 registros en orden alfabético ascendente consumiendo el API /api/medicamentos. Para los efectos adversos se consume el API /api/efectosAdversos, con los mismos parámetros 100 registros en orden alfabético ascendente. Por último, para las alertas sanitarias se consume el API /api/alertas-sanitarias limitado a las últimas 20 alertas registradas, lo cual muestra las alertas más actuales emitidas por la COFEPRIS.

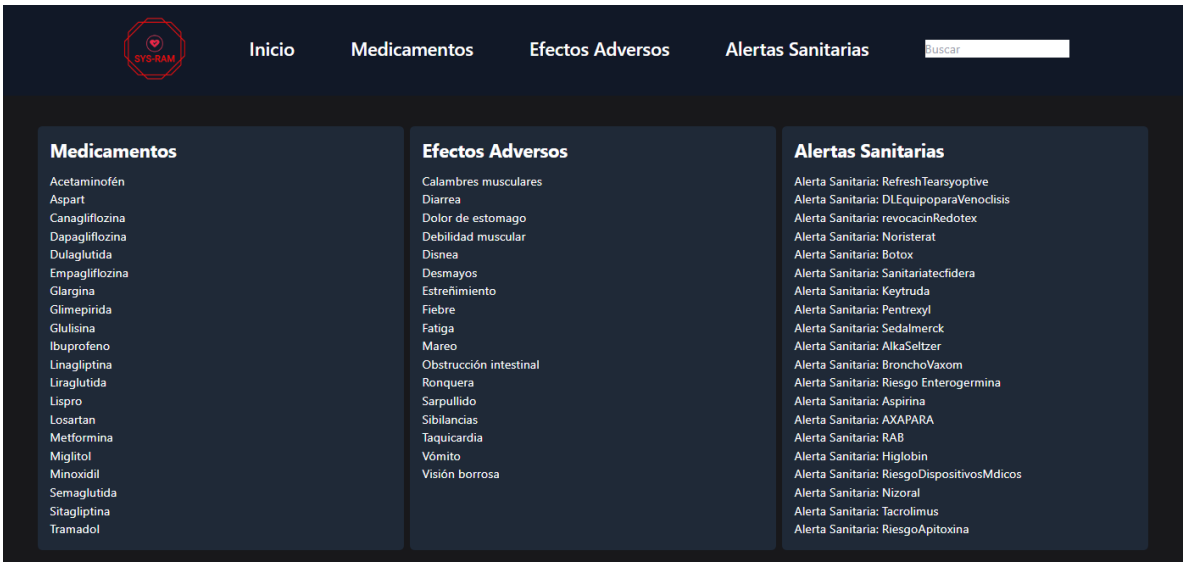


Figura 27 Página principal de sistema Sys-RAM.

Para la vista de medicamentos, la cual se aprecia en la Figura 28, de lado izquierdo se muestra un componente con las letras iniciales de los medicamentos, el cual seleccionarlo carga automáticamente los medicamentos con esa letra inicial, se muestra el nombre del medicamento y el inicio de sus condiciones de uso con un límite de 150 caracteres.

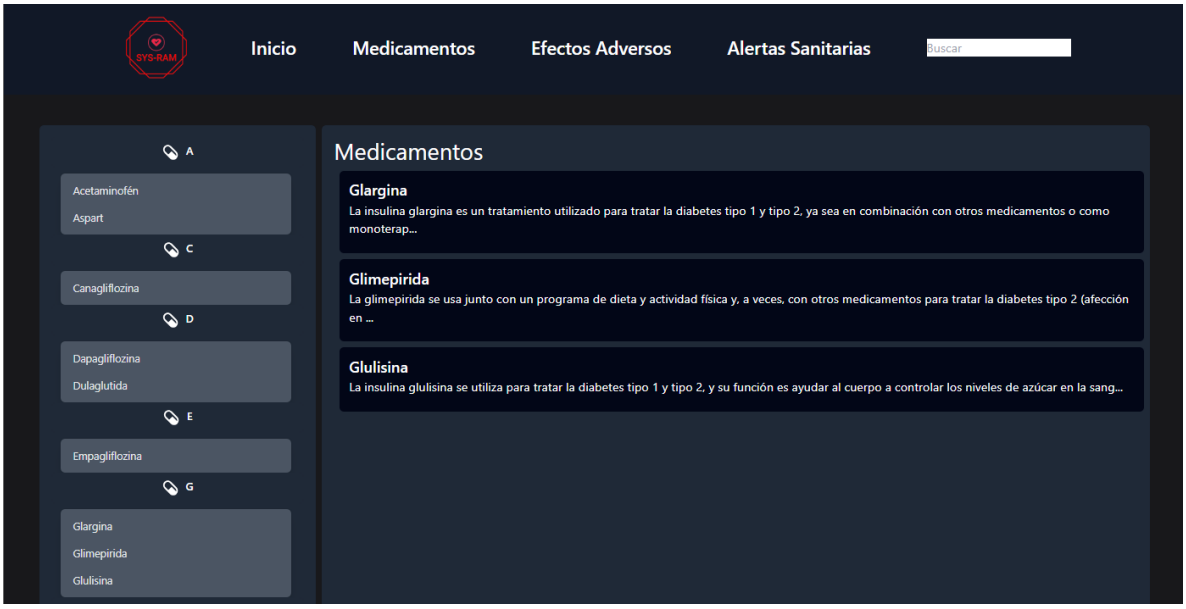


Figura 28 Vista del listado de medicamentos.

Así mismo al seleccionar el detalle del medicamento, se carga la información del mismo, en apartados separados y desplegables para minimizar el espacio en pantalla y el usuario pueda decidir que apartado es de su interés, tal y como se muestra en la Figura 29.

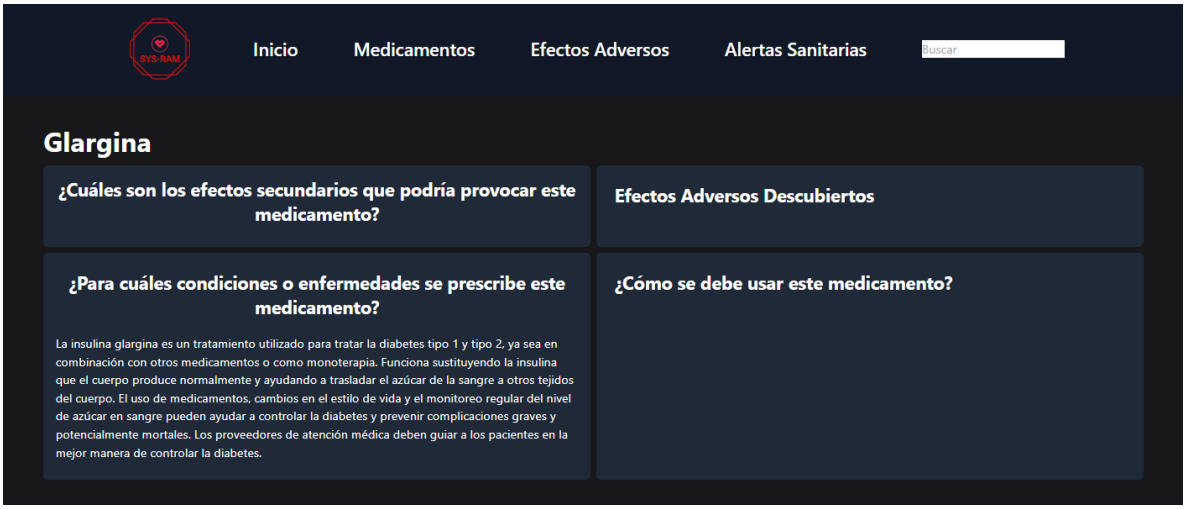


Figura 29 Vista de destalle de medicamentos.

Para el apartado de los Efectos Adversos se reutilizan los componentes brindando así una sensación de cohesión en la aplicación web y cargando la información correspondiente, así mismo se muestran los efectos adversos ordenados de manera alfabética y su descripción limitada a 150 caracteres, tal y como se aprecia en la Figura 30.

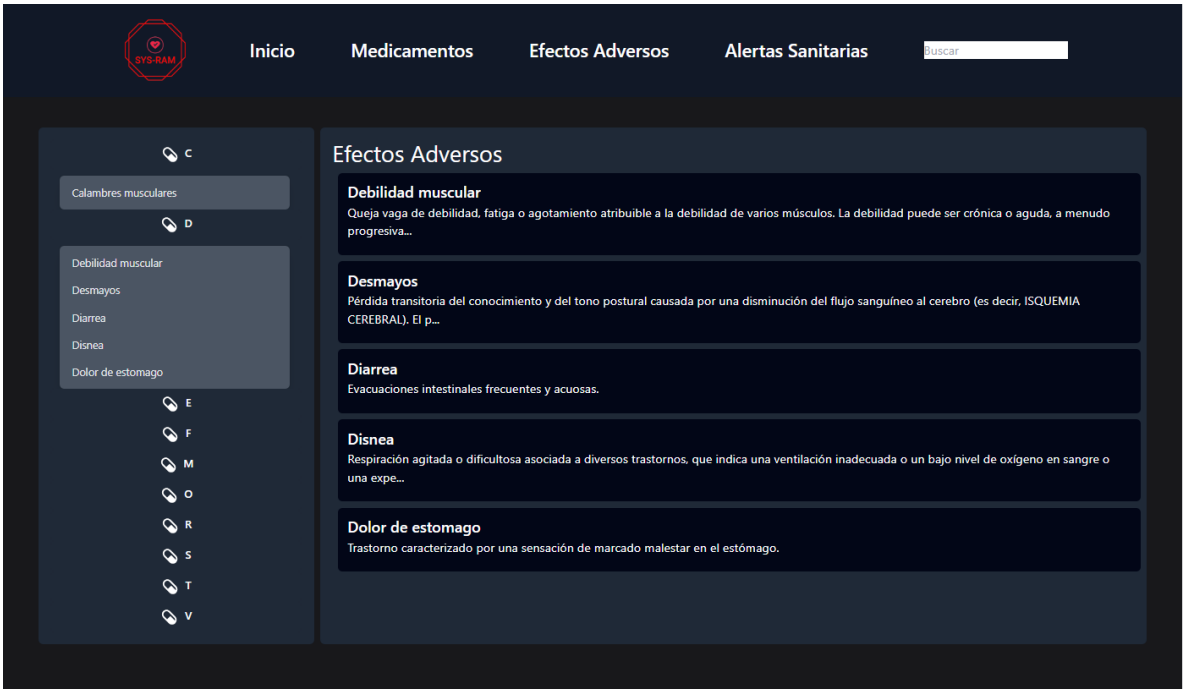


Figura 30 Vista de listado de efectos adversos.

En la Figura 31 se puede observar que, al seleccionar el detalle de uno de los efectos adversos, se muestra el nombre del mismo, así como la descripción formal completa.

Figura 31 Vista de detalle de efectos adversos.

Para el apartado de las alertas sanitarias se muestra el listado de las ultimas 20 alertas emitidas por la COFEPRIS así mismo se muestra el PDF el cual detalla toda la información de la alerta y que puede ser consultada por cualquier usuario (ver Figura 32).

Figura 32 Vista de alertas sanitarias y visualización de detalle de las mismas.

Vistas de aplicación móvil

La aplicación móvil se diseñó con base en los mockups creados en el apartado anterior, en la Figura 33 se aprecia la primer pantalla de la aplicación móvil la cual muestra la lista de medicamentos ordenados de manera alfabética para poder consultar su información a detalle, así como permitir la búsqueda y filtrado por nombre del medicamento.

Figura 33 Vista de listado de medicamentos en aplicación móvil.

Al seleccionar un medicamento se muestran sus condiciones de uso, su forma de uso y finalmente los efectos adversos tanto los reportados como los descubiertos por el desarrollo de esta investigación, esto se puede apreciar en la Figura 34.

Figura 34 Vista de detalle de medicamento de la aplicación móvil.

Otro apartado que es de vital importancia y que brinda información relevante a los usuarios es poder consultar las alertas sanitarias, esto se aprecia en la Figura 35, donde se listan las alertas sanitarias ordenadas por fecha de emisión, así mismo es posible buscar entre las alertas sanitarias existentes.

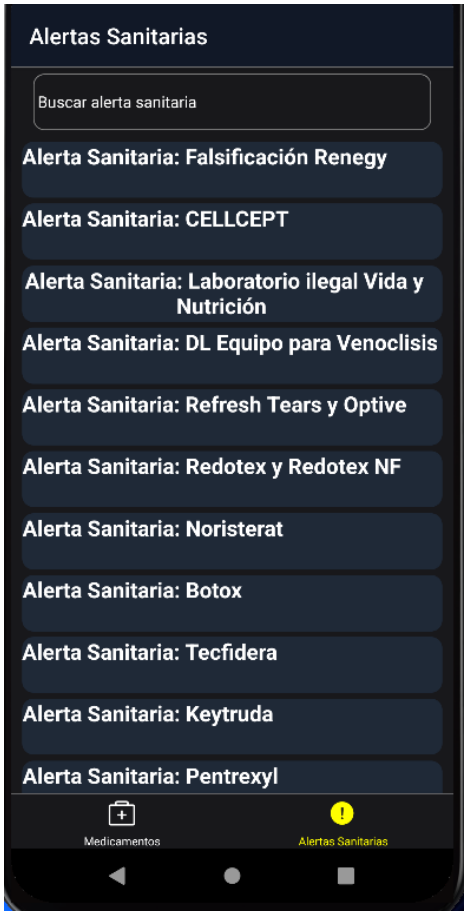


Figura 35 Vista de listado de alertas sanitarias en aplicación móvil.

Finalmente, el detalle de las alertas sanitarias puede ser consultado al seleccionar cualquiera de ellas, esto consulta su PDF asociado y da la posibilidad de visualizarlo sin salir de la aplicación, esto es apreciable en la Figura 36.

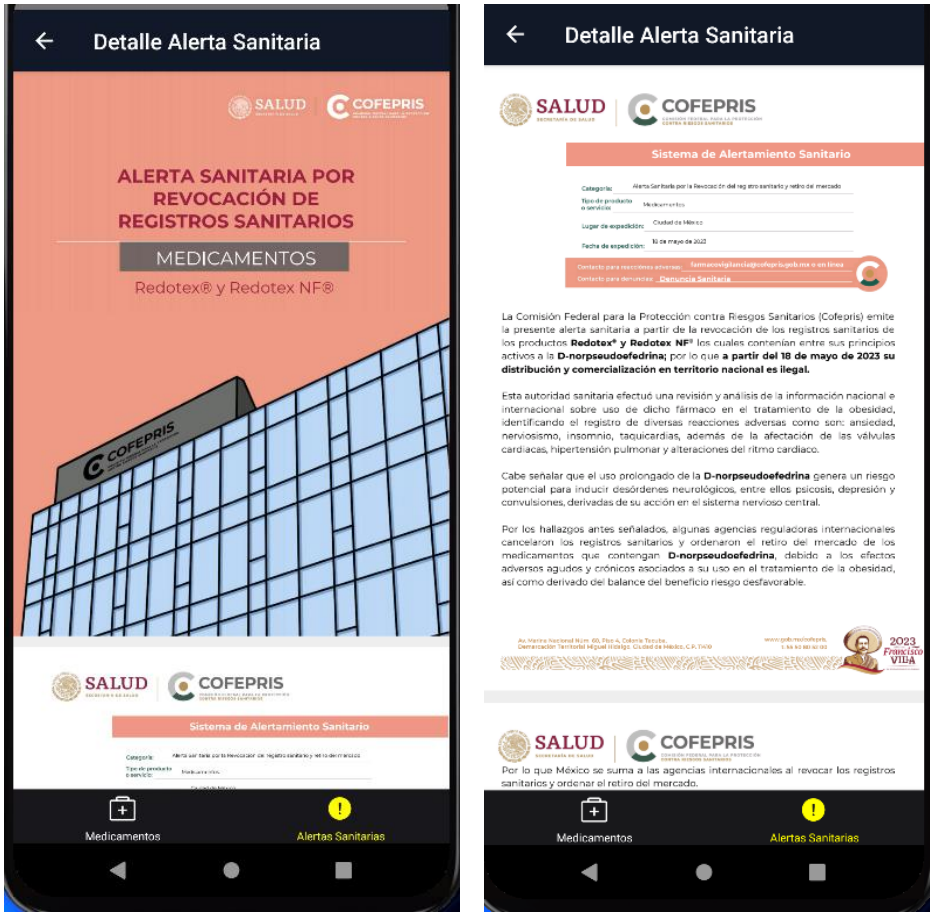


Figura 36 Vista de detalle de alerta sanitaria.

3.3 Diseño de base de datos

Debido a que la base de datos elegida para el desarrollo de este proyecto es una base de datos no relacional esta puede ser representada de diversas maneras, tales como elementos de clave y valor, columnas y familias, grafos, y documentos. El diseño se hará a través de documentos ya que esta es la forma en la que MongoDB almacena los datos.

Colección alertas_sanitarias

```
{
"_id": "index",
"enlace": "string",
"nombre": "string",
"fecha": "string",
"updated_at": ISODate (),
"created_at": ISODate ()
}
```

Colección efectos_adversos

```
{
"_id": ObjectId (),
"nombre": "string",
"descripcion": "string",
"medicamentosRelacionados": [],
"codigo": "string",
"letra": "string",
"updated_at”: ISODate (),
"created_at”: ISODate ()
}
```

Colección medicamentos

```
{
"_id": ObjectId (),
"nombre": "string",
"condicionUso": {
      "titulo": "string",
      "contenido": ["string", "string", …]
},
"comoUsar": {
      "titulo": "string",
      "contenido": ["string", "string", …]
},
"efectosAdversos": {
      "titulo": "string",
      "contenido": ["string", "string", …]
},
"efectosAdversosDescubiertos": ["string", "string", …],
"updated_at": ISODate (),
"created_at": ISODate (),
}
```

3.4 Arquitectura

La arquitectura del sistema desarrollado se basa en una estructura en capas distinguiéndose 3 capas principales, que son la capa de aplicación, capa de dominio y capa de datos. En la capa de aplicación se tienen los clientes que son la aplicación web desarrollada en React y la aplicación móvil desarrollada en React Native que se conectan a la lógica o capa de dominio bajo un esquema cliente-servidor en la cual se realizan todos los procesos de comunicación, manejo de peticiones y comunicación entre la aplicación y la base de datos del proyecto, así como base de datos externas. Así mismo la capa de dominio tiene acceso al modelo desarrollado para la clasificación de nuevos comentarios y el registro de efectos adversos descubiertos y relacionados a medicamentos registrados, todo esto se puede ver representado en la Figura 37.

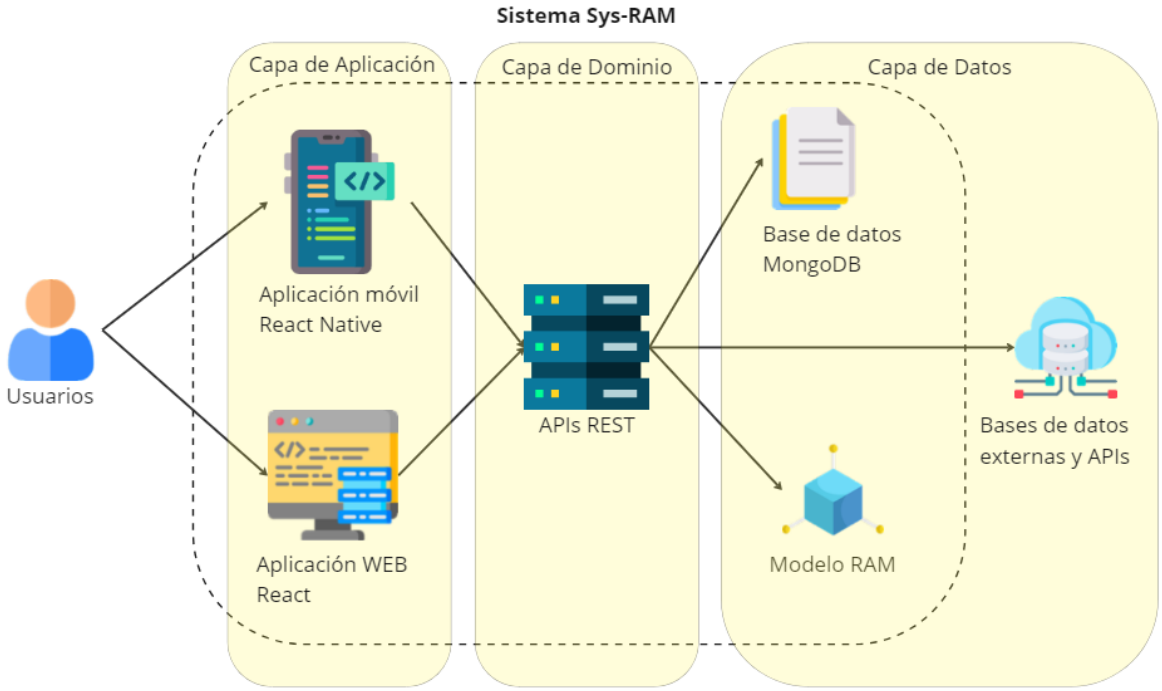


Figura 37 Diseño de la arquitectura del sistema Sys-RAM.

CAPITULO IV RESULTADOS

4.1 Validación del modelo desarrollado

En el apartado de la evaluación del modelo dentro de la metodología CRISP-DM se mostraron resultados sobre los experimentos realizados con el corpus generado debido a que era un paso necesario para comprobar que el modelo creado cumplía con los objetivos de la metodología. En esta sección se expandirá la descripción de esos resultados mostrados en la Tablta 10. Por qué la red neuronal RAM obtuvo los mejores resultados puede deberse a la arquitectura de la propia red neuronal la cual le permite tener distintas capas de atención y recordar palabras que tienen cierta lejanía con el efecto adverso. También se puede notar que RAM obtuvo un valor de exactitud de 86%, y muy cerca el algoritmo SVM. Esto puede deberse a que el algoritmo SVM contenía más información para lograr una mayor exactitud. Así mismo, la red neuronal IAN obtuvo resultados muy cercanos a la red neuronal RAM sin embargo estos fueron inferiores tanto a nivel de Valor-F con un 76% y a nivel de exactitud con un 84%.

Como se puede evidenciar en los resultados, hablando en términos de Valor-F estos son menores con respecto a lo presente en el estado del arte ejemplos de esto se pueden apreciar en (I. S. Alimova & Tutubalina, 2020; Sakhovskiy & Tutubalina, 2022) donde se obtuvieron Valores-F de 96.4% y 79.9% respectivamente. Estos resultados pueden deberse a distintas razones como lo puede ser el tamaño del corpus utilizado en los experimentos respecto a los existentes en idioma inglés es más pequeño, esto debido a que no hay corpus existentes de efectos adversos en idioma español formados por información de redes sociales. Otro punto para tomar en cuenta es que el idioma español representa una mayor complejidad dada la gramática más flexible del mismo comparado con el idioma inglés. Finalmente, los diccionarios de efectos adversos, medicamentos y vectores de palabras incrustadas aún no son los suficientemente maduros como si lo son en inglés.

4.2 Caso de estudio

4.2.1 Análisis de medicamento Metformina para el tratamiento de diabetes, muestra de efectos adversos reportados y descubiertos

Para el desarrollo de este caso de estudio se utilizará el sistema desarrollado Sys-RAM en su versión de aplicación web, con el fin de evaluar su utilidad e información brindada a aquellos usuarios que quieran conocer los efectos adversos relacionados específicamente al medicamento metformina, dado a que es uno de los medicamentos más recetados para el tratamiento de la diabetes. Entre las preguntas que un usuario puede realizarse sobre un medicamento se desprenden las siguientes:

- ¿Qué efectos adversos tiene un medicamento?
- Si un efecto adverso no aparece o no ha sido reportado, pero lo padezco es ¿es posible que se un efecto adverso no reportado?
- ¿Cuál es la forma de utilizar el medicamento correctamente?
- ¿Para qué situaciones se prescribe un medicamento?

De acuerdo con las preguntas anteriores se procederá a realizar los pasos para poder las respuesta a estas. Entrando al sistema se muestra la pantalla de inicio con los medicamentos, entre los cuales se puede visualizar la metformina, tal y como se muestra en la Figura 38. Así mismo, es posible visualizar el apartado Medicamentos en la barra de navegación desde la cual se pueden visualizar todos los medicamentos existentes, para el caso de la Metformina esta se encuentra en la letra inicial M y al seleccionarla se despliegan los medicamentos con esa inicial además de mostrar una vista previa de su condición de uso tal y como se muestra en la Figura 39. Se puede ingresar desde cualquiera de las dos opciones mostradas, para este caso de estudio se ingresó desde el apartado medicamentos.

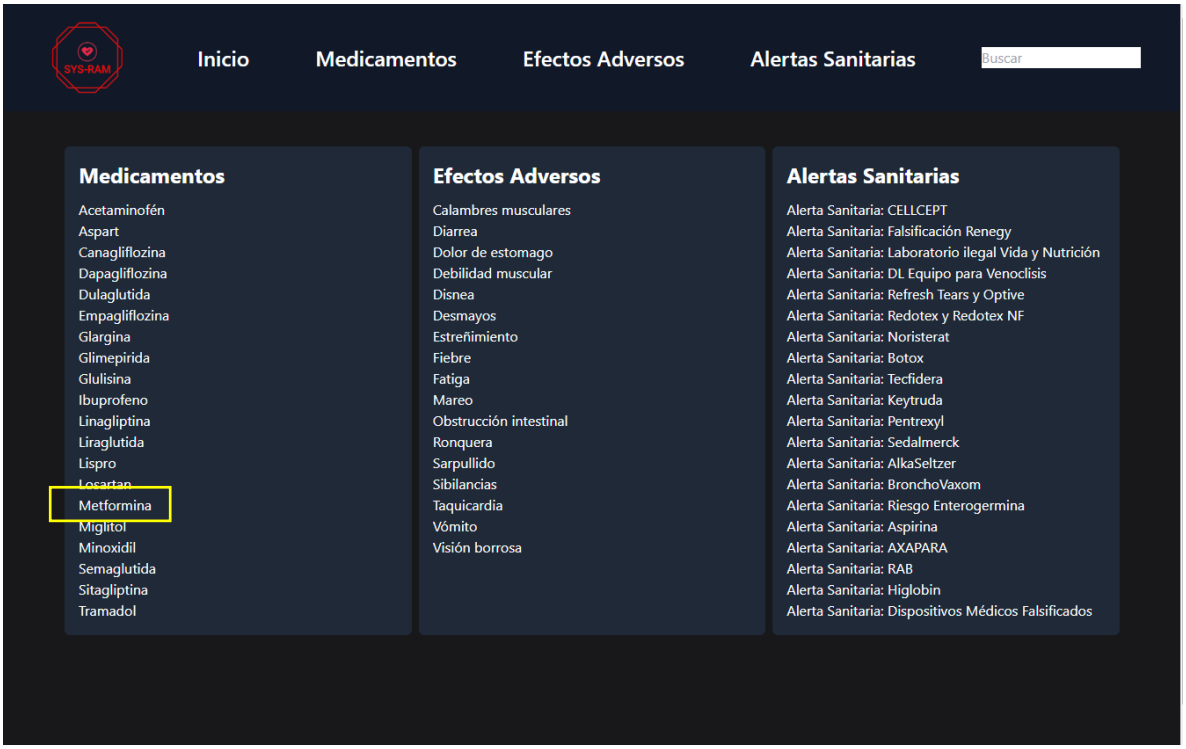


Figura 38 Vista principal y selección de medicamento Metformina para caso de estudio.

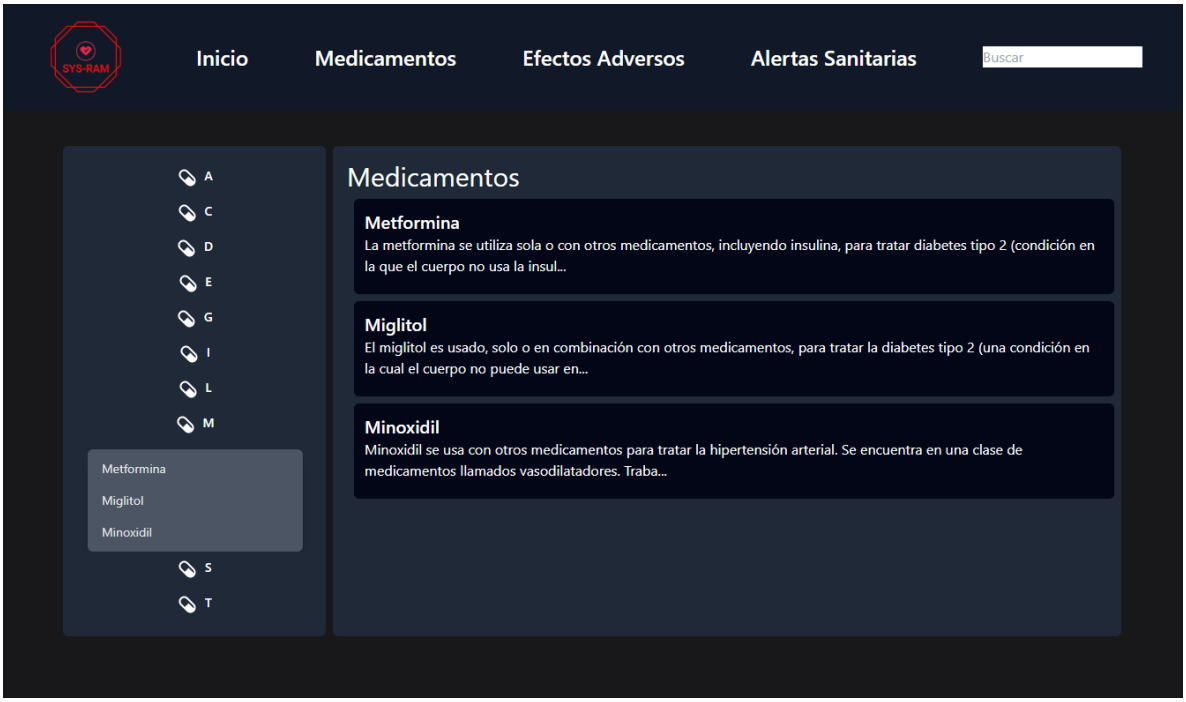


Figura 39 Selección y visualización de medicamentos con inicial M.

Al ingresar a la vista del medicamento Metformina se muestran las secciones de manera minimizada para que el usuario decida qué información es de su interés tal

y como se visualiza en la Figura 40. Para este caso se mostrarán los efectos adversos reportados oficialmente, los cuales tienen un color dependiendo de su nivel de peligrosidad para la salud de las personas donde los colores verde y amarillo son efectos adversos leves o moderados además de ser los más comunes, mientras que los marcados con color naranja o rojo son efectos adversos significativos o graves y se recomienda hablar con su doctor para que le indique al usuario que acciones realizar. Así mismo se muestran los efectos adversos descubiertos mediante el modelo creado anteriormente, estos se obtuvieron y registraron de aquellos comentarios que tenían presencia de efectos adversos, esto se muestra en la Figura 41. Un punto para tomar en cuenta es que los efectos adversos descubiertos pueden coincidir con los efectos adversos reportados de manera oficial, también su color es azul ya que hasta que no sean clasificados por un experto de la salud no se puede afirmar que un efecto adverso tenga cierto nivel de impacto en la salud del usuario.

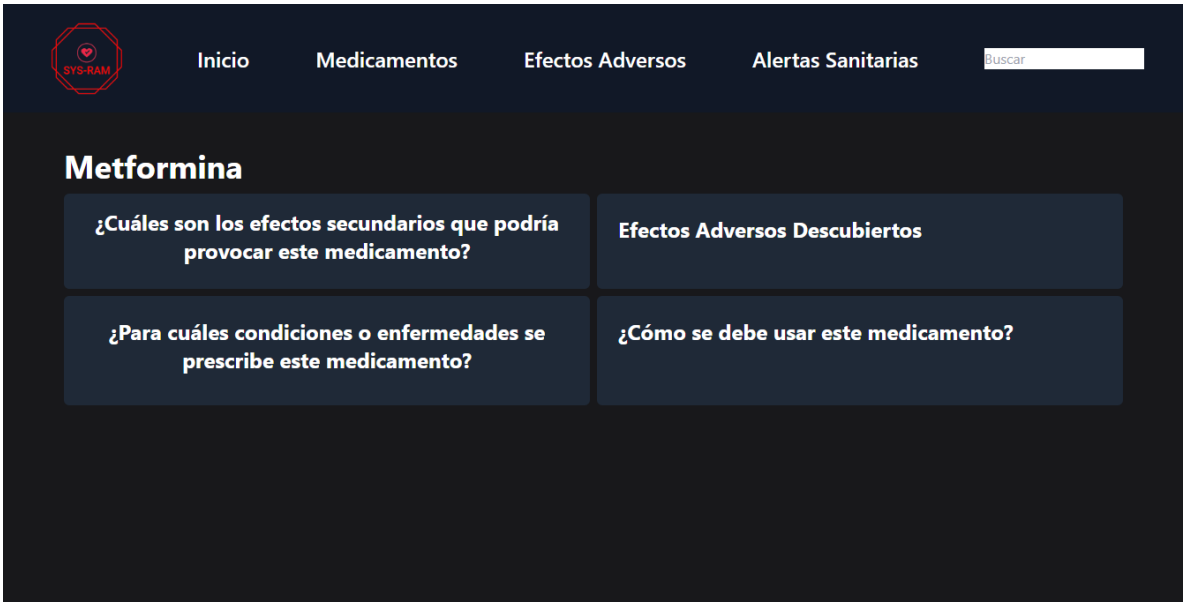


Figura 40 Visualización inicial de medicamento Metformina con secciones ocultas.

Figura 41 Visualización de efectos adversos registrados y descubiertos del medicamento Metformina.

Otra información que también es relevante para los usuario es la o las condiciones de prescripción del medicamento, así como la forma de usar, ya que un mismo medicamento puede tener distintas presentaciones y el importante poder consultar esta información, esto se evidencia en la Figura 42, en donde se describe la prescripción de la Metformina y sus diversas maneras de uso lo cual brinda información de vital importante para una persona que lleve un tratamiento de Metformina o para aquellos usuarios que cuidan de personas y tomen el medicamento.

Figura 42 Visualización de información extra sobre condiciones de prescripción y forma de uso de medicamento Metformina.

Finalmente, si el usuario requiere una mayor descripción de los efectos adversos puede acceder mediante la página de inicio o mediante la barra de navegación, y buscar el efecto adverso de su interés tal y como se muestra en la Figura 38 y en la Figura 43. En este caso un efecto adverso muy común que causa la Metformina es el Dolor de Estómago, en la Figura 44 se puede apreciar la descripción detalla del mismo.

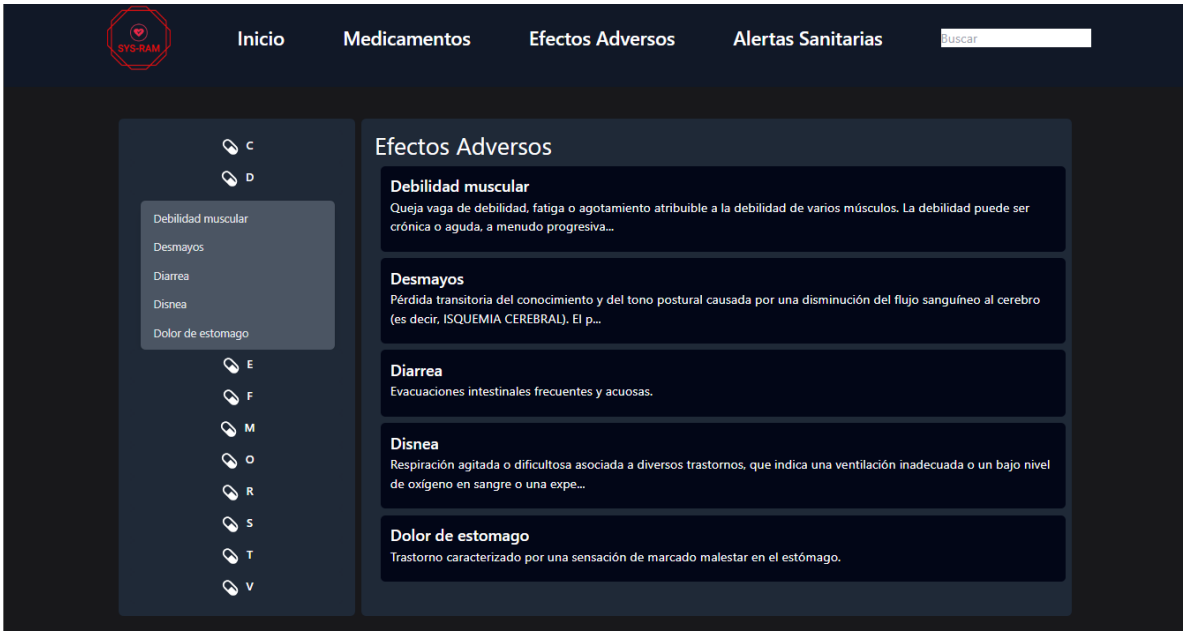


Figura 43 Visualización de lista de efectos adversos y búsqueda de aquellos asociados al medicamento Metformina.

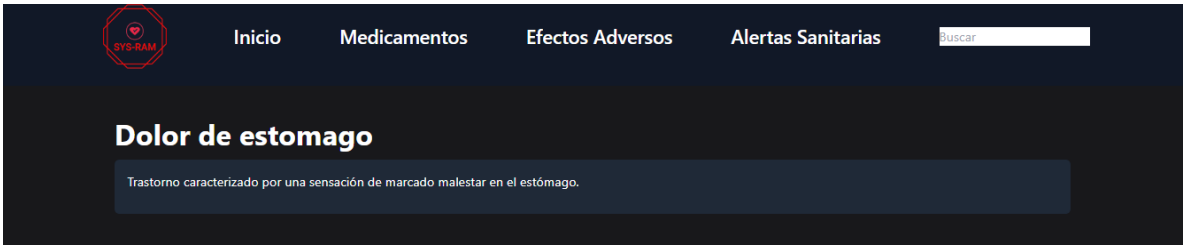


Figura 44 Detalle de efecto adverso asociado al medicamento Metformina.

4.2.2 Consulta de alertas sanitarias

Para realizar la prueba de este caso de estudio en Sys-RAM se procedió a ejecutar de manera manual la tarea programa para que recolectara las alertas sanitarias disponibles hasta el día 13 de junio de 2023, esta obtención de alertas se realizó de manera exitosa tal y como se puede apreciar en la Figura 45.

```
λ php artisan health-alert:cron
Cron Job, ejecutandose obtención de alertas ejecutandose a las: 2023-06-13 07:55:34
Cron Job, finalizó la obtención de alertas ejecutandose a las: 2023-06-13 07:55:35
```

Figura 45 Ejecución manual de tarea programada para comprobación de funcionamiento.

Estas alertas pueden ser consultadas desde la página de inicio como se muestra en la Figura 38 y desde la barra de navegación, al estar ordenadas por la fecha de emisión de la alerta siempre se mostrarán las últimas alertas emitidas por la COFEPRIS, así un usuario puede visualizar cualquier alerta que sea de su interés. Seleccionando la primer alerta que puede ser la de mayor interés para el usuario se

muestra que el medicamento Cellcept de 500mg está circulando de manera ilegal, además esta alerta fue emitida el día 12 de junio de 2023 mostrando así que el margen de emisión de la alerta y la disponibilidad de la misma para la consulta de los usuario es bastante oportuna, esto se puede apreciar en la Figura 46.

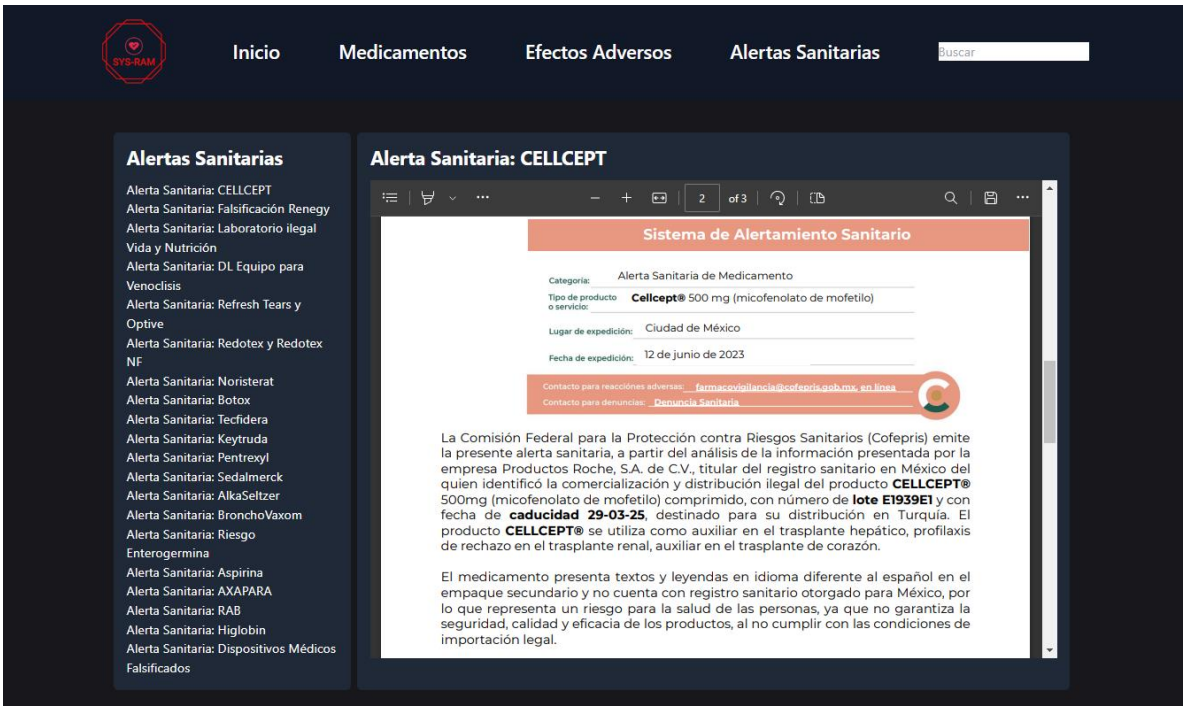


Figura 46 Vista de detalle de alerta sanitaria más reciente emitida por la COFEPRIS.

4.3 Comprobación de la hipótesis

Para la comprobación de la hipótesis se consultó con la misma doctora quien facilitó los medicamentos presentes en la Tabla 7. Se le proporcionaron los efectos adversos obtenidos de redes sociales del medicamento Metformina, el cual fue el caso de estudio, los efectos adversos son los siguientes: Daño renal, pérdida de apetito, sensibilidad estomacal, dolor abdominal, dolor de cabeza, migraña, náuseas, acidez estomacal, deposiciones liquidas, distensión abdominal, episodios depresivos, vértigo y mareos, sensación de sabor metálico, pérdida de peso e irritabilidad. Con esos efectos adversos, la doctora confirmo que son RAM que se presentan comúnmente en personas que toman metformina, sobre todo aquellos relacionados al sistema gastrointestinal, esto confirma la hipótesis de esta investigación ya que el sistema desarrollado es capaz de detectar los efectos adversos causados por medicamentos presentes en redes sociales.

CAPITULO V CONCLUSIONES

5.1 Conclusiones

Con el desarrollo de este trabajo de tesis, se permitió la creación de un modelo capaz de clasificar comentarios con presencia o ausencia de efectos adversos con un porcentaje mayor a lo que está presente actualmente en el estado del arte siendo un problema que ha sido poco explorado en el idioma español, obteniendo un *Accuracy* del 86%. Así mismo la creación de un corpus en idioma español que recopile los comentarios de redes sociales es un aporte de gran impacto al no existir en la literatura un corpus con estas características actualmente.

Así mismo el desarrollo del sistema Sys-RAM que integra información recopilada del modelo desarrollado, así como información recopilada de medicamentos, alertas sanitarias y efectos adversos mediante técnicas de web scraping para la presentación de manera amigable, intuitiva y oportuna a los usuarios tiene como resultado el cumplimiento del objetivo general de esta investigación. Así mismo la arquitectura por capas diseñada como parte de la metodología de desarrollo permitió la integración de distintas fuentes de información, el desarrollo de la lógica de la capa del dominio, permitiendo así la creación de una aplicación web y una aplicación móvil, haciéndola adaptable, modular y mantenible en el tiempo. Aunado a lo anterior, la arquitectura permite extense a la integración de nuevos módulos, el cambio de desarrollo o incluso la creación de una aplicación de escritorio para hospitales o médicos que la requieran para consultas constantes.

Finalmente, es importante recalcar que Sys-RAM puede seguir creciendo y siendo alimentado por más información de medicamentos, más información de efectos adversos y nuevas alertas sanitarias emitidas por la COFEPRIS, teniendo así un sistema robusto que se irá actualizando a lo largo del tiempo y siempre tendrá información integra y de calidad para los usuarios.

5.2 Recomendaciones y trabajo a futuro

Existen algunas recomendaciones de acuerdo con lo desarrollado en este trabajo de tesis, ya que a pesar de que se cumplieron con los objetivos tanto general como los específicos, se identificaron ciertas mejoras tanto a nivel de modelo como del sistema Sys-RAM.

Una de las recomendaciones es el recolectar más información sobre más medicamentos relacionados con la hipertensión y diabetes, así como el registro o asociación de marcas de laboratorios asociadas a la producción de ciertos

medicamentos para cubrir un rango más amplio de los mismos. Así mismo, se recomienda el ampliar tanto los medicamentos como los efectos adversos relacionados de acuerdo a otras enfermedades crónico-degenerativas tales como artritis, obesidad, osteoartritis, EPOC y cáncer, al ser enfermedades muy presentes en la sociedad mexicana. Otro punto para mejorar es la relación de los efectos adversos con los medicamentos y mostrar esta relación desde el efecto adverso para dar a conocer a los usuarios que medicamento puede causar cierto efecto adverso de manera específica.

Como trabajo a futuro se propone el ampliar el corpus mediante una recolección más extensa de datos de redes sociales, ampliado con la búsqueda de más medicamentos, así como la recolección de comentarios de blogs especializados en la salud esto con el fin de mejorar el modelo desarrollado. Otro punto como trabajo a futuro es el habilitar a los usuarios la función de realizar comentarios dentro de Sys-RAM para que el modelo vaya aprendiendo de manera automática con los comentarios realzados. Y finalmente, el ampliar la información de los medicamentos, mostrando datos técnicos de los mismos que pueden ser de gran utilidad para usuarios especializados en el área de la salud.

REFERENCIAS

Alimova, I. S., & Tutubalina, E. V. (2020). Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models. *Programming and Computer Software 2019 45:8*, *45*(8), 439–447. https://doi.org/10.1134/S0361768819080024

Alimova, I., & Tutubalina, E. (2018). Automated detection of adverse drug reactions from social media posts with machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10716 LNCS*, 3–15. https://doi.org/10.1007/978-3-319-73013-4_1/COVER

Alonso-Zárate, J., & Casas Roma, J. (2021). *Bases de datos no relacionales*.

Basiri, M. E., Abdar, M., Cifci, M. A., Nemati, S., & Acharya, U. R. (2020). A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. *Knowledge-Based Systems*, *198*, 105949. https://doi.org/10.1016/J.KNOSYS.2020.105949

Camps Paré, R., Casillas Santillán, L. A., Costal Costa, D., Gilbert Ginestà, M., Martín Escofet, C., & Pérez Mora, O. (2005). *Bases de datos*. Eureca Media. https://www.uoc.edu/pdf/masters/oficiales/img/913.pdf

Cañete, J. (2019). *Compilation of Large Spanish Unannotated Corpora*. https://doi.org/10.5281/ZENODO.3247731

Cardellino, C. (2016, March). *Spanish Billion Word Corpus and Embeddings*. https://crscardellino.ar/SBWCE/

Chapman, A. B., Peterson, K. S., Alba, P. R., DuVall, S. L., & Patterson, O. v. (2019a). Detecting Adverse Drug Events with Rapidly Trained Classification Models. *Drug Safety*, *42*(1), 147–156. https://doi.org/10.1007/S40264-018-0763-Y/TABLES/12

Chapman, A. B., Peterson, K. S., Alba, P. R., DuVall, S. L., & Patterson, O. V. (2019b). Detecting Adverse Drug Events with Rapidly Trained Classification Models. *Drug Safety*, *42*(1), 147–156. https://doi.org/10.1007/S40264-018-0763-Y/TABLES/12

Dolores Barrientos, Y., Villar García, M. G., González Calderón, D. E., Portilla Luja, M. de las M., Villaseñor Contreras, M., MALDONADO REYES ANA AURORA, /, Serrano Barquín, H. P., Serrano Barquín, C., Uribe Rosas, C., Mora Cantellano,

M. del P. A., Torres Fragoso, A. M., Espinosa Hernández, M. del C., Berrelleza Rendón, A., Rivera Castillo, S. G., Molina González, M. N., Morales González, C. G., Pichardo Beltrán, I. K., Sosa Compeán, L. B., Luna Rodríguez, S. A., … Hernández Romero, Y. (2022). *Diseño para grupos vulnerables en tiempos de crisis*. http://ri.uaemex.mx/handle/20.500.11799/112066

Edo-Osagie, O., de La Iglesia, B., Lake, I., & Edeghere, O. (2020). A scoping review of the use of Twitter for public health research. *Computers in Biology and Medicine*, *122*, 103770. https://doi.org/10.1016/J.COMPBIOMED.2020.103770

Gräßer, F., Malberg, H., Kallumadi, S., & Zaunseder, S. (2018). Aspect-Based sentiment analysis of drug reviews applying cross-Domain and cross-Data learning. *ACM International Conference Proceeding Series*, *2018-April*, 121–125. https://doi.org/10.1145/3194658.3194677

Gupta, S., Pawar, S., Ramrakhiyani, N., Palshikar, G. K., & Varma, V. (2018). Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction. *BMC Bioinformatics*, *19*(8), 1–7. https://doi.org/10.1186/S12859-018-2192-4/TABLES/2

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour 2021 5:4*, *5*(4), 529–538. https://doi.org/10.1038/s41562-021-01079-8

Lavertu, A., Hamamsy, T., & Altman, R. B. (2021a). Quantifying the Severity of Adverse Drug Reactions Using Social Media: Network Analysis. *J Med Internet Res 2021;23(10):E27714 Https://Www.Jmir.Org/2021/10/E27714*, *23*(10), e27714. https://doi.org/10.2196/27714

Lavertu, A., Hamamsy, T., & Altman, R. B. (2021b). Quantifying the Severity of Adverse Drug Reactions Using Social Media: Network Analysis. *J Med Internet Res 2021;23(10):E27714 Https://Www.Jmir.Org/2021/10/E27714*, *23*(10), e27714. https://doi.org/10.2196/27714

Liu, Y., Shi, J., & Chen, Y. (2018). Patient-centered and experience-aware mining for effective adverse drug reaction discovery in online health forums. *Journal of the Association for Information Science and Technology*, *69*(2), 215–228. https://doi.org/10.1002/ASI.23929

Ma, D., Li, S., Zhang, X., & Wang, H. (n.d.). *Interactive Attention Networks for Aspect-Level Sentiment Classification*. Retrieved May 9, 2023, from http://alt.qcri.org/semeval2014/task4/

Martín Mateos, J. L., & Ruiz Reina, F. J. (2013). *Procesamiento del lenguaje natural*. https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf

Olston, C., & Najork, M. (2010). Web Crawling. *Foundations and Trends R in Information Retrieval*, *4*(3), 175–246. https://doi.org/10.1561/1500000017

Organización Mundial de la Salud. (2019). *OMS INDICADORES DE FARMACOVIGILANCIA: UN MANUAL PRÁCTICO PARA LA EVALUACIÓN DE LOS SISTEMAS DE FARMACOVIGILANCIA*. https://apps.who.int/iris/bitstream/handle/10665/325851/9789243508252-spa.pdf?ua=1

Pauli, P. A. (2019). *Análisis de sentimiento: Comparación de algoritmos predictivos y métodos utilizando un lexicon español* [Instituto Tecnológico de Buenos Aires]. https://ri.itba.edu.ar/server/api/core/bitstreams/db2a9097-b8f4-4205-8048-8f9fdc76cd66/content

Peng, C., Zhongqian, S., Lidong, B., & Yang, W. (n.d.). *Recurrent Attention Network on Memory for Aspect Sentiment Analysis*. 452–461.

Piñeiro Gómez, J. M. (2014). *Diseño de bases de datos relacionales*. Paraninfo.

Porto Arceo, J. Á. (2019). Reacciones adversas a medicamentos. Generalidades. Criterios de derivación. *SEICAP*, 285–295. https://www.aeped.es/sites/default/files/documentos/20_ra_medicamentos_generalidades.pdf

Sakhovskiy, A., & Tutubalina, E. (2022). Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics*, *135*, 104182. https://doi.org/10.1016/J.JBI.2022.104182

Sarker, A., Belousov, M., Friedrichs, J., Hakala, K., Kiritchenko, S., Mehryary, F., Han, S., Tran, T., Rios, A., Kavuluru, R., de Bruijn, B., Ginter, F., Mahata, D., Mohammad, S. M., Nenadic, G., & Gonzalez-Hernandez, G. (2018). Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, *25*(10), 1274–1283. https://doi.org/10.1093/JAMIA/OCY114

Stenhouse, N. V. (2017). *HabScraper: herramienta automatizada para la extracción de datos con web scraping*. https://dspace.uib.es/xmlui/bitstream/handle/11201/151095/Memoria_EPSU1195.pdf?sequence=1

Suárez-Paniagua, V., & Segura-Bedmar, I. (2018). Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction. *BMC Bioinformatics*, *19*(8), 39–47. https://doi.org/10.1186/S12859-018-2195-1/TABLES/6

Torruella, J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e Informática. Nuevas Tecnologías En Los Estudios Filológicos. Barcelona: Seminario de Filología e Informática*, 45–77. http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

Vilar, S., Friedman, C., & Hripcsak, G. (2018a). Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media. *Briefings in Bioinformatics*, *19*(5), 863–877. https://doi.org/10.1093/BIB/BBX010

Vilar, S., Friedman, C., & Hripcsak, G. (2018b). Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media. *Briefings in Bioinformatics*, *19*(5), 863–877. https://doi.org/10.1093/BIB/BBX010

Wang, C. S., Lin, P. J., Cheng, C. L., Tai, S. H., Yang, Y. H. K., & Chiang, J. H. (2019). Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model. *J Med Internet Res 2019;21(2):E11016 Https://Www.Jmir.Org/2019/2/E11016*, *21*(2), e11016. https://doi.org/10.2196/11016

Zhao, B. (2017). Web Scraping. *Encyclopedia of Big Data*, 1–3. https://doi.org/10.1007/978-3-319-32001-4_483-1

Zunic, A., Corcoran, P., & Spasic, I. (2020). Sentiment Analysis in Health and Well-Being: Systematic Review. *JMIR Med Inform 2020;8(1):E16023 Https://Medinform.Jmir.Org/2020/1/E16023*, *8*(1), e16023. https://doi.org/10.2196/16023