DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCION I.- TESIS

TRABAJO PROFESIONAL

"SISTEMA DE ANÁLISIS DE OPINIONES EN ESPAÑOL DE MEDICAMENTOS PARA ENFERMEDADES CRÓNICO-DEGENERATIVAS BASADO EN TÉCNICAS DE WEB SCRAPING"

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN
SISTEMAS COMPUTACIONALES

PRESENTA:
I.S.C KARINA CASTRO PÉREZ

DIRECTOR DE TESIS:
DR. JOSÉ LUIS SÁNCHEZ CERVANTES

CO-DIRECTOR DE TESIS:
DRA. MARÍA DEL PILAR SALAS ZÁRATE

ORIZABA, VER. MÉXICO NOVIEMBRE 2020

"2020, Año de Leona Vicario, Benemérita Madre de la Patria"

Orizaba, Veracruz, 13/11/2020
Dependencia: División de Estudios de Posgrado e Investigación
Asunto: Autorización de Impresión
OPCION: I

C. KARINA CASTRO PÉREZ
Candidato a Grado de Maestro en:
SISTEMAS COMPUTACIONALES
P R E S E N T E.-

De acuerdo con el Reglamento de Titulación vigente de los Centros de Enseñanza Técnica Superior, dependiente de la Dirección General de Institutos Tecnológicos de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

" SISTEMA DE ANÁLISIS DE OPINIONES EN ESPAÑOL DE MEDICAMENTOS PARA ENFERMEDADES CRÓNICO-DEGENERATIVAS BASADO EN TÉCNICAS DE WEB SCRAPING"

Comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE ORIZABA
DIV. DE EST. DE POSGRADO E INVEST.

DR. MARIO L. ARRIOJA RODRÍGUEZ
JEFE DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN

FECHA: **21/07/2020**

ASUNTO: Revisión del Trabajo Escrito

C. MARIO LEONCIO ARRIOJA RODRIGUEZ
JEFE DE LA DIVISION DE ESTUDIOS
DE POSGRADO E INVESTIGACION
P R E S E N T E.

Los que suscriben miembros del jurado, han realizado la revisión de la Tesis del (la) C.:

KARINA CASTRO PÉREZ
No. DE CONTROL: **M13011098**

La cual lleva el título de:

"Sistema de análisis de opiniones en español de medicamentos para enfermedades crónico-degenerativas basado en técnicas de Web Scraping"

Y concluyen que se acepta.

A T E N T A M E N T E

PRESIDENTE:	DR. JOSÉ LUIS SÁNCHEZ CERVANTES	_____
SECRETARIO:	M.S.C. LUIS ÁNGEL REYES HERNÁNDEZ	_____
VOCAL:	DRA. LISBETH RODRÍGUEZ MAZAHUA	_____
VOCAL SUP.:	DRA. MA. DEL PILAR SALAS ZÁRATE	_____

EGRSADO (A) DE LA MAESTRIA EN **SISTEMAS COMPUTACIONALES**

OPCION: **1 Tesis**

Índice

Índice de tablas

Índice de figuras

Resumen

En la actualidad, la minería de opiniones es un tema de gran interés e importancia en diversas áreas, porque permite resumir y analizar la información contenida en la Web, para proporcionar datos útiles para la toma de decisiones. La minería de opiniones, entre sus diversas áreas de aplicación, destaca por su contribución en el contexto del cuidado de la salud. Por ello, el objetivo de la presente tesis es realizar un sistema de minería de opiniones de medicamentos para enfermedades crónico-degenerativas, en el idioma español, en el cual, se implementen técnicas de Web Scraping y el Procesamiento del Lenguaje Natural (PLN). Con base en la literatura analizada se encontró que existen trabajos que abordan el área del cuidado de la salud usando la minería de opiniones, sin embargo, gran parte de este trabajo se centra en analizar la información contenida en las redes sociales, especialmente en Twitter, en los idiomas como el inglés, chino y árabe, dejando una gran oportunidad de investigación en fuentes de datos como foros y video blogs, en el idioma español. El enfoque propuesto, a través del sistema, tiene como objetivo en primer lugar, proporcionar a los especialistas información que ayude a agilizar el proceso de identificación y selección de los medicamentos que prescriben, también, conocer los efectos adversos que tienen otros pacientes con los medicamentos prescritos para su condición, para identificar rápidamente los efectos adversos de sus pacientes, permitiéndoles dedicar más tiempo al examen físico y evitar así complicaciones adicionales a la enfermedad, lo que se traduce en una mayor calidad de atención a los pacientes. En segundo lugar, proporcionar a los pacientes información para saber lo que piensan los demás sobre los medicamentos que utilizan, y con la validación de los expertos, se les facilite a los pacientes identificar los comentarios que no representan un peligro para su salud y los que fomentan la automedicación. Los resultados obtenidos son alentadores ya que se obtuvo una *Precision* del 87% y 82% para opiniones positivas y negativas, respectivamente, y un *Recall* del 94% para opiniones positivas y 65% para opiniones negativas. Además, el *Accuracy* del modelo usado para la detección de polaridad de un 86%.

Abstract

Nowadays, opinion mining is a topic of great interest and importance in several areas because it allows for the summary and analysis of the information in the Web, providing useful data for decision making. Opinion mining, among its various areas of application, stands out for its contribution to healthcare. Therefore, this thesis's objective is to develop a system for opinion mining of drugs for chronic-degenerative diseases in the Spanish language, in which Web Scraping and Natural Language Processing (NLP) techniques are implemented. Based on the literature analyzed, we found that there are works that address the area of healthcare using opinion mining. However, much of this work focuses on analyzing the information contained in social networks, especially Twitter, in languages such as English, Chinese, and Arabic, leaving an excellent opportunity for research in data sources such as forums and video blogs, in the language in Spanish. The proposed approach, through the system, aims firstly, to provide specialists with information to help speed up the process of identifying and selecting the medicines they prescribe, secondly, to know the adverse effects that other patients have with the medicines prescribed for their condition, to quickly identify the adverse effects of their patients, and to spend more time on the physical examination to avoid additional complications to the disease, which translates into a higher quality of care for patients. Secondly, to provide patients with information to find out what others think about the medicines they use, and with expert validation, patients can identify comments that do not pose a danger to their health and encourage self-medication. The results obtained are encouraging since a Precision of 87% and 82% were obtained for positive and negative opinions, respectively, and a Recall of 94% for positive and 65% for negative opinions. In addition, the Accuracy of the model used for the polarity detection was 86%.

Introducción

En los últimos años, la aplicación de la minería de opiniones aumentó por el crecimiento de las redes sociales y los sitios Web, fuentes de datos con grandes volúmenes de información, sin embargo, la revisión manual es poco factible. Por esta razón, es necesario aplicar las técnicas de Web Scraping y de minería de opiniones, dos técnicas primarias que ayudan a obtener y resumir los datos. La minería de opiniones, tiene diversas áreas de aplicación, es muy visible su uso en el comercio electrónico, por ejemplo, cuando alguien califica un producto y conforme a ello se sacan estadísticas para conocer si a las personas les gusta, de igual forma en las redes sociales cuando se publican comentarios acerca de un tema, servicio, o persona, por mencionar algunos.

Sin embargo, la minería de opiniones destaca en el ámbito del cuidado de la salud por los beneficios que proporciona en la toma de decisiones, no obstante, se requieren de estudios que aborden el cuidado de la salud en el idioma español, que permita conocer de forma fácil y rápida los medicamentos prescritos para las enfermedades crónico-degenerativas, tales como diabetes mellitus tipo 2, hipertensión, asma, Alzheimer, entre otros.

Por consiguiente, el presente proyecto aborda tal problemática aplicando la minería de opiniones dirigida hacia el análisis para la obtención de la polaridad de los medicamentos prescritos para las enfermedades crónico-degenerativas, identificando parámetros clave, como la efectividad y efectos adversos, tomando como fuente de datos: foros, sitios Web especializados y video blogs. Cabe destacar que no se incluye como fuente de datos las redes sociales de Twitter y Facebook, pues con base en la literatura analizada se encontró que gran parte del trabajo existente se centra en realizar minería de opiniones en las redes sociales.

El presente trabajo describe el enfoque propuesto para dar solución al problema planteado, el cual se realizó bajo la metodología UWE (*Unified Modeling Language Web Engineering*, Ingeniería Web del Lenguaje Unificado de Modelado). Además, haciendo uso de técnicas y herramientas, como lo son: el desarrollo de un *Crawler* con la ayuda de la API (*Application Programming Interface*, Interfaz de Programación de Aplicaciones) Jsoup y JDOM (*Java Document Object Model*, Java Modelo de Objeto del Documento)para realizar el *Scraping*; una

bolsa de palabras con nombres de medicamentos que se obtienen de una Ontología de dominio médico, haciendo uso del editor Protégé para poblar dicha ontología y la API de Apache Jena para desarrollar de la bolsa de palabras, así mismo el uso de la API de Stanford CoreNLP para la obtención de la clasificación de la polaridad.

De modo que, este trabajo de tesis de maestría se compone por cinco capítulos, el capítulo uno aborda el marco teórico con conceptos clave para la comprensión, así como los objetos del proyecto a efectuar y la justificación; el capítulo dos se centra en el estado de la práctica y un análisis comparativo de los estudios, enfoques y tecnologías empleadas; en el capítulo tres se describe la metodología, así como la arquitectura diseñada para dar solución al problema planteado y la metodología de la ontología que se emplea; el capítulo cuatro se compone de dos casos de estudio que prueban el funcionamiento de SentiScrap, así como los resultados del experimento realizado; finalmente, el capítulo cinco contiene las conclusiones y recomendaciones para trabajo futuro.

Capítulo 1. Antecedentes

A continuación, se explican los conceptos esenciales, descritos para la mejor comprensión del trabajo presentado, además, se aborda una descripción detallada del problema planteado y con ello los objetivos generales y específicos que dan solución al planteamiento de la presente tesis.

1.1 Marco teórico

1.1.1 Web Scraping

El Web *Scraping* se conceptualiza como la construcción de un agente informático o un programa que tiene por objetivo descargar, analizar y organizar datos de la Web de manera automática para un propósito en específico [1]. Este tipo de programas indaga y solicita los datos a partir de palabras clave o de datos de sentimientos utilizando "robots" de software, los cuales efectúan una simulación de búsqueda realizada por humanos, los que los hace ser anónimos y difíciles de detectar [2].

Existen cuantiosos casos de uso diferentes en los que se requiere el uso del Web *Scraping*, gran parte de los casos se centran en el hecho de que la información que se encuentra en Internet a menudo se encuentra de una forma dispar, sin embargo, al aplicar las técnicas del Web *Scraping* y reunir los datos en un solo conjunto resultan ser muy valiosos, ya que se emplean para llevar a cabo análisis, extraer la información necesaria y con ello obtener datos estadísticos [3].

1.1.1.1 Web Crawling

Web *Crawling* es una técnica del Web *Scraping*, también conocida como s*pider* (araña) o robots, los cuales se caracterizan por ser programas informáticos para colectar automáticamente páginas Web, dado a que en la actualidad existe una gran cantidad de información relevante que se encuentra dispersa en la Web. Es así que Web *Crawler* surge por la necesidad del uso de rastreadores que ayuden a extraer información de la Web ya que es una entidad dinámica que crece a un ritmo acelerado [4].

El Web *Crawler* tiene un comportamiento similar a la navegación que realizan los usuarios, es decir, de primera instancia se especifica el conjunto de URL (*Uniform Resource Locator*, Localizador Uniforme de Recursos) a las cuales tiene que acceder para posteriormente ingresar

a cada página y descargar el contenido, efectuando un proceso repetitivo hasta terminar con todas las URL asignadas. Al término de su ejecución devuelve al usuario la información solicitada, para llevar a cabo el análisis de los datos. Generalmente, los motores de búsqueda como lo es Google emplea *Crawler* para obtener contenido de una URL y examinar cada página Web para encontrar más enlaces, sin embargo, ejecutan varios procesos en paralelo para llevar a cabo dicha encomienda, pero a una mayor velocidad [5], [6].

1.1.1.2 Web Information Extraction

La extracción de información Web es una técnica del Web *Scraping*, el cual identifica información estructurada y semiestructurada que interesa a los usuarios. El objetivo principal es obtener texto de las páginas Web, imágenes, audios, video y otros medios.

En la actualidad, los métodos de extracción de información Web tienen una variedad de categorías, las cuales se listan a continuación:

1. **Dividido según el objeto de la extracción de información Web**: desestructurado, extracción de texto, extracción de texto semiestructurado y extracción de texto estructurado;
2. **Dividido en función del grado de automatización**: extracción artificial de información, extracción semiautomática de información, y extracción automática de información;
3. **Dividido de acuerdo con el sistema de extracción de información Web existente y el principio de implementación del modelo**: extracción de información basada en el aprendizaje inductivo, extracción de información basada en HTML (*HyperText Markup Language*, Lenguaje de Marcado de Hipertexto), extracción de información basada en ontologías, extracción de información basada en el Procesamiento del Lenguaje Natural (PLN) y extracción de información basada en consultas Web [7].

1.1.2 Minería de opiniones

La minería de opiniones o también llamado el análisis de sentimientos es el proceso de extracción de información subjetiva a partir de contenido generado por el usuario, también, usa el PLN y la inteligencia computacional [8].

Este campo de estudio analiza las opiniones, sentimientos, valoraciones, emociones y actitudes de los usuarios hacia entidades como productos, servicios, organizaciones, individuos, eventos, temas y sus atributos. Existen muchos nombres y tareas que se diferencian como: análisis de sentimientos, extracción de opiniones, extracción de sentimientos, análisis de subjetividad, análisis de afectos, análisis de emociones, análisis de minería, entre otros.

Sin embargo, están todos bajo el concepto del análisis de sentimiento o la minería de opinión. Mientras que, el término de análisis de sentimientos es más común para la industria, por otra parte, el uso del término de minería de opiniones es frecuentemente usado en el ámbito académico [9].

1.1.2.1 Detección de la polaridad

Este tipo de análisis es una de las principales tareas de la minería de opiniones, el cual busca establecer el estado emocional de las opiniones contenidas en un texto determinado, con palabras tales como, “bien”, “triste” o “feliz”, que equivale a neutro, negativo y positivo, respectivamente. Además, lo polaridad tiene relación con los niveles de la minería de opiniones, por ello, se aplica a documentos, párrafos u oraciones particulares que contienen información subjetiva, como reseñas de artículos o publicaciones en blogs [10], [11].

1.1.3 Ontología

Una ontología es una descripción explícita de un dominio que puntualiza un concepto describiendo sus propiedades, atributos y restricciones de una manera organizada para limitar la complejidad. Por lo tanto, la extracción de la información que se basa en una ontología permite la identificación y la clasificación empleando un sistema basado en reglas y un análisis de nivel más profundo [12].

De igual forma brinda una definición de reglas y axiomas, que permiten el uso de mecanismos informales para así descubrir datos invisibles. Pero, una de las cosas más importantes que aporta una ontología es la interconexión de múltiples fuentes de datos [13].

1.1.3.1 Lenguaje de Otología Web

OWL (*Web Ontology Language*, Lenguaje de Ontología Web) se diseñó con el propósito de ser empleado en las aplicaciones que requieren del procesamiento de contenido, por ello, se usa

para representar claramente el significado de los términos en un vocabulario, así como las relaciones existentes entre dichos términos.

El Lenguaje de Ontología Web tiene una gran capacidad para la interpretación Web que la soportada por XML (*Extensible Markup Language*, Lenguaje de Marcado Extensible), RDF (*Resource Description Framework*, Marco de Descripción de Recurso) y RDF- *Schema*, es así, que proporciona un vocabulario adicional con una semántica formal.

De igual forma, OWL provee tres sub-lenguajes expresivos y diseñados para implementarlos en comunidades específicas.

A continuación, se listan los tres sub-lenguajes:

- OWL Lite: Presenta compatibilidad con los usuarios que requieren de una jerarquía para la clasificación y las restricciones simples.
- OWL DL: El cual es compatible con los usuarios que precisan de una máxima expresividad, pero a la vez que conservan la integridad computacional y decidibilidad.
- OWL Full: Dirigido a los usuarios que necesitan de una máxima expresividad y libertad sintáctica de RDF, el cual es un modelo de datos para los objetos y las relaciones entre ellos [14].

1.1.3.2. Lenguaje de Otología Web 2

Es un lenguaje de representación de conocimiento, el cual se diseñó para formular, intercambiar y razonar con conocimiento sobre un dominio de interés. Las ontologías OWL 2 proporcionan clases, propiedades, individuos y valores de datos, además, se almacenan como documentos Web semánticos. Por lo cual, las ontologías OWL 2 se usan junto con la información escrita en RDF.

OWL 2 es un lenguaje declarativo, porque describe el estado de una cosa de una manera lógica. Mientras que las herramientas apropiadas se usan para inferir más información sobre el estado de las cosas. De modo que, la forma en que se realizan las inferencias algorítmicamente no forma parte del documento OWL, sino que depende de las implementaciones específicas. Por

lo cual, solo las implementaciones que cumplan con esta semántica se considerarán acordes con OWL 2.

Por lo cual, OWL 2 no es un lenguaje de esquema porque no proporciona medios elaborados para prescribir cómo se estructura sintácticamente un documento. Tampoco es un *Framework* de base de datos a pesar de que los documentos OWL almacenan información, sin embargo, OWL 2 traza una cierta analogía entre la información terminológica y el contenido de la base de datos [15].

1.1.4 Bolsa de Palabras

BOW (*Bag of Words*, Bolsa de Palabras) es un modelo ampliamente usado en el domino del Procesamiento de Lenguaje Natural porque se basa en la idea de que la frecuencia de aparición de una palabra en un texto sirve como una medida del significado que tiene dicha palabra [16].

El modelo de Bolsa de Palabras pasa por alto la gramática y el orden, por ejemplo, las siguientes dos frases son semánticamente diferentes: "*Es una niña bonita, ¿verdad?* y, *ella no es bonita, ¿es niña?*". De modo que el modelo BOW utiliza las palabras que son significativas de un lenguaje y que representan las particularidades de un documento para mantenerlas en un conjunto que representa una bolsa [17].

1.1.5 Normalización del texto

La normalización de texto o también conocida como el pre-procesamiento de texto es una técnica que consta de tres tareas principales: eliminación de caracteres especiales que no proporcionan información relevante, corrección de errores ortográficos y la sustitución de las abreviaturas [18].

De la misma forma, la tokenización es una tarea importante dentro de la normalización de texto, la cual consiste en fragmentar un texto a los que se les denomina tokens. De modo que, un token corresponde a una unidad de estructura gramática, tal como lo es un párrafo, una oración o una palabra de un texto en concreto. Sin embargo, el texto al que se le aplica la tokenización, pasa por un procesamiento mediante la aplicación de algoritmos de normalización de texto [19], [20] y [21].

1.1.6 Corpus

El tratamiento de los datos a través de un corpus ha demostrado ser eficaz para encontrar la solución a problemas existentes de la lingüística, por ello, el creciente auge de dicha disciplina. Si bien el corpus no tiene relación directa con la informática, se ha hecho una adaptación de un corpus informatizado, para facilitar la organización y la explotación de grandes cantidades de información.

El corpus informatizado es una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con el propósito de ser tratados mediante procesos informáticos, que son útiles para reflejar el comportamiento de una o más lenguas [22].

1.1.7 Base de Datos

Conjunto de datos estructurados que emplea numerosos archivos los cuales almacenan y organizan datos [23]. La información contenida tiene características en común, es almacenada en un equipo de cómputo para que los sistemas de software, controlados por los usuarios, que tengan autorización accedan a dicha base para consultar la información. Además, una Base de Datos se conforma con "*un esquema lógico junto con las instancias de los datos operativos que dicho esquema organiza*" [24].

1.1.7.1 Sistema Gestor de Base de Datos

El Sistema Gestor de Base de Datos (SGBD) es un software que provee una interfaz entre los datos y sistemas de software externos. El SGBD es el encargado de garantizar el correcto, eficiente, íntegro y seguro acceso y almacenamiento de los datos. Además de permitir definir, construir y manipular una Base de Datos [25].

De modo que al definir una Base de Datos se especifican los tipos de datos, las estructuras y las restricciones. Por otra parte, construye una Base de Datos para almacenar información y manipular se refiere a hacer las operaciones de consultar y actualizar los datos [26].

1.1.8 Foro

El foro es un servicio que ofrece el Internet, como una forma de comunicación que permite el intercambio de opiniones sobre un tema en específico *"... es una forma de discusión colectiva, de análisis y de intercambio de ideas"*. Los foros se crearon para exponer temas de la actualidad,

abordar asuntos de interés social y para propagar información relevante que permite a los usuarios reflexionar, complementar, debatir, argumentar o crear ideas de un tema de interés [27].

1.1.9 Video blog

Un video blog o un Vlog se constituye sobre la base de clips de video ordenada cronológicamente. Que pueden ser registros propios o de otros usuarios. Los comentarios son una parte fundamental en la estructura de un blogs o video blogs ya que se trata de una herramienta que promueve el dialogo [28].

En la actualidad existen diversas plataformas de video como lo es YouTube y Vimeo, las cuales permite a las personas subir contenido de cualquier tipo y compartir los videos a nivel mundial. En dichas plataformas se encuentran canales que van de un contenido específico a diverso, en los cuales se habla de diversos temas incluyendo libros, educación, salud, viajes, juegos, entre muchos otros, en donde expresan una opinión negativa o positiva respecto a la experiencia que vivieron.

1.2 Planteamiento del problema

En la actualidad, existe una gran cantidad de información no estructurada en el contexto médico y de la salud contenido en blogs, video blogs, redes sociales y en sitios Web especializados. Sin embargo, la información representa un papel trascendental en el proceso de cuidado de la salud, porque permite a los pacientes y a profesionales de la salud conocer las experiencias y opiniones de la comunidad en relación a aspectos tales como ejercicio, alimentación, medicamentos, entre otros.

No obstante, la revisión manual de la información resulta inviable dado a la gran cantidad de datos que se genera en tiempo real. Por ello, es necesaria la aplicación de técnicas de minería de opiniones que permitan resumir toda la información, así como proporcionar el entorno y las herramientas necesarias para el intercambio de conocimientos y el apoyo mutuo.

De forma que la minería de opiniones es una de las áreas de investigación más activas en el PLN y también es ampliamente estudiada en minería de datos, minería Web y recuperación de la información. La creciente importancia de la minería de opiniones coincide con el crecimiento de las redes sociales y de los blogs.

De modo que los sistemas de minería de opiniones se están aplicando en diversos ámbitos comerciales y sociales, ya que las opiniones son fundamentales para casi todas las actividades humanas y son factores clave que influyen en el comportamiento del ser humano.

En este trabajo se propone un enfoque para el análisis de opiniones de medicamentos para enfermedades crónico-degenerativas en el idioma español, basado en la técnica de Web *Scraping* para obtener opiniones desde sitios Web especializados, blogs y video blogs.

1.3 Objetivo general y específicos

A continuación, se describe a detalle el objetivo general y los objetivos específicos.

1.3.1 Objetivo general

Desarrollar un sistema de software que permita conocer y analizar las opiniones proporcionadas por los pacientes en idioma español, hacia medicamentos para el cuidado de enfermedades crónico-degenerativas a través de técnicas de minería de opiniones y Web *Scraping*.

1.3.2 Objetivos específicos

- Analizar el trabajo relacionado para conocer técnicas sobre minería de opiniones y Web *Scraping*.
- Identificar sitios Web especializados en opiniones sobre medicamentos para enfermedades crónico-degenerativas, en el idioma español.
- Analizar API's como herramientas para efectuar el Web *Scraping* y la minería de opiniones.
- Diseñar una arquitectura que dirija el proceso para efectuar el análisis de polaridad de opiniones sobre medicamentos relacionados a enfermedades crónico-degenerativas.
- Desarrollar un *crawler* para reducir tiempo y esfuerzo en la búsqueda de información en español sobre medicamentos de enfermedades crónico-degenerativas
- Efectuar un módulo de extracción de datos de foros especializados y de extracción de subtítulos de video blogs sobre medicamentos, con respecto a factores de precio, efectos adversos, efectividad, entre otros.
- Desarrollar un módulo de normalización de las opiniones extraídas.
- Diseñar un corpus, para ayudar a la detección de polaridad de los comentarios obtenidos.

- Realizar una Bolsa de Palabras de dominio médico a través de una ontología.
- Realizar un módulo para la aplicación del PLN y de la minería de opiniones.
- Proporcionar de una manera factible y rápida los medicamentos con mejor reputación dentro de la comunidad.
- Efectuar al menos un caso de estudio en el dominio de la salud específicamente para enfermedades crónico-degenerativas.

1.4 Justificación

Se requieren herramientas que ayuden a los pacientes que padecen enfermedades crónico-degenerativas, tales como: diabetes mellitus tipo 2, hipertensión, hepatitis, Alzheimer, entre otros, a tener acceso a información confiable sobre los medicamentos prescritos para los tratamientos de las enfermedades mencionadas. Con base en lo anterior, este trabajo de tesis lleva a cabo el diseño de un enfoque híbrido basado en la semántica y el aprendizaje automático supervisado para un sistema que implementa la minería de opiniones y PLN, así como el Web *Scraping* para obtener opiniones desde sitios Web especializados, blogs y video blogs que mencionen medicamentos para las enfermedades crónico-degenerativas, en concreto, diabetes, hipertensión y hepatitis, en el idioma español.

Lo que resulta en una herramienta Web que permite a los pacientes conocer de manera fácil y rápida los medicamentos con mejor reputación dentro de la comunidad ante la gran cantidad de marcas y nombres existentes en el mercado, y facilitar a los especialistas de la salud información para ayudar a agilizar el proceso de identificación y selección de los medicamentos que prescriben, además, conocer los efectos adversos que tienen otros pacientes con los medicamentos prescritos para su condición, para identificar rápidamente los efectos adversos de sus pacientes, permitiéndoles dedicar más tiempo al examen físico y evitar así complicaciones adicionales a la enfermedad, para una mayor calidad de atención a los pacientes.

Capítulo 2. Estado de la práctica

A continuación, se presenta un análisis detallado del trabajo relacionado, así como una comparativa sobre características específicas que se incluye en cada uno de los trabajos analizados, así mismo se describen las diferencias y similitudes del proyecto efectuado con el trabajo existente. Además, en este capítulo se muestran las herramientas y metodologías idóneas para la solución del problema plateando en la presente tesis.

2.1 Trabajos relacionados

Lee *et al.* [29] elaboraron un análisis que describe la dificultad de encontrar los efectos adversos en ensayos clínicos que realizan las organizaciones farmacéuticas y la falta de vigilancia requerida en el mercado para identificar efectos secundarios no descubiertos previamente. Por ello, los autores examinaron los modelos de aprendizaje profundo, donde identificaron que el recurso es costoso dada la escasez de tweets de ADE (*Adverse Drug Events*, Eventos Adversos a Medicamentos) de Twitter, sin embargo, los enfoques semi-supervisados son una alternativa conveniente y un recurso potencial para detectar ADE en tiempo real en las publicaciones de Twitter. En este contexto, construyeron modelos basados en CNN (*Convolutional Neural Network*, Red Neuronal Convolucional) utilizando una arquitectura semi-supervisada para la clasificación de ADE en tweets, haciendo uso selectivo de una variedad de datos sin etiquetar. Además, los experimentos realizados de tweets aleatorios con: 1) Nombres médicos; 2) Condiciones de salud; 3) Oraciones de artículos científicos en la literatura médica y Wikipedia, y 4) Oraciones simuladas relacionadas con la salud creadas a partir de léxicos, demostraron que la clasificación de ADE supera los modelos de clasificación supervisada, un +9.9% medida de F1 y los modelos de supervisado de vanguardia con una precisión de +14.58%.

Alayba *et al.* [30] identificaron la existencia de una gran cantidad de tweets generados cada minuto en el idioma árabe, que incluyen temas sobre servicios de salud donde frecuentemente aparecen en tendencias de la red social de Twitter. Por tal motivo, se aplicaron cuatro pasos para la obtención de un nuevo conjunto de datos en árabe, para fines de minería de opinión, tales como: 1) Recuperación de datos; 2) Filtrado de datos; 3) Pre-procesamiento del texto en árabe, mediante la eliminación de datos no deseados, la eliminación de algunas palabras y el texto no

relacionados, así como la normalización del texto, y 4) Aplicación del aprendizaje automático al conjunto de datos recopilado. Por otra parte, se efectuaron experimentos utilizando redes neuronales profundas y algoritmos de aprendizaje automático, tales como, *Naïve Bayes*, LR *(Logistic Regression*, Regresión Logística), SVM (*Support Vector Machines*, Máquina de Soporte Vectorial), DNN (*Deep Neural Network*, Red Neuronal Profunda) y CNN (*Convolutional Neural Network*, Red Neuronal Convolucional) donde se obtuvieron resultados con una precisión aproximadamente entre el 85% y el 91%, además, se encontró que los mejores clasificadores son de SVM utilizando la clasificación de vectores de soporte lineal y el descenso de gradiente estocástico.

Cavalcanti y Prudêncio [31] centraron su análisis en la minería de opinión para la revisión de medicamentos, en la que los pacientes expresan sus experiencias y opiniones sobre tratamientos o medicamentos. Así mismo, identificaron que la farmacovigilancia beneficia a los fabricantes de medicamentos porque las reacciones adversas particulares a un medicamento se rastrean rápidamente desde las publicaciones en redes sociales o foros públicos. Por ello, los autores propusieron la adopción de un método lingüístico para extraer aspectos en las revisiones de medicamentos basado en las rutas de dependencia en el árbol sintáctico de las revisiones, para extraer pares de opinión, es decir, un término de aspecto y un término de opinión. De igual forma, la iniciativa de los autores incluyó un algoritmo de aprendizaje supervisado para la clasificación de los pares de opiniones en uno de los cuatro tipos de aspectos: Condición, ADR (*Adverse Drug Reactions*, Reacción Adversa a Medicamentos), dosis o efectividad. Finalmente, llevaron a cabo experimentos con conjuntos de datos relacionados con tres enfermedades diferentes como el Trastorno por Déficit de Atención e Hiperactividad (TDAH), SIDA y ansiedad, los resultados revelaron una mejora en el rendimiento para extraer aspectos relevantes en comparación con los métodos de referencia, donde los valores más altos de *F-Measure* se observaron para todos los conjuntos de datos, además, se encontró que la solución propuesta se adapta fácilmente a otros idiomas, ya que no requiere datos etiquetados.

Los pacientes que consumen fármacos a menudo buscan historias de pacientes como ellos en Internet que, por lo general, encuentran entre sus amigos y familiares, sin embargo, pocos estudios que indagan sobre el impacto de las redes sociales en los pacientes demuestran que,

para algunos problemas de salud el apoyo comunitario en línea resulta en un efecto positivo. En este contexto, Gopalakrishnan y Ramaswamy [32] propusieron un método de extracción de opinión que se centra en predecir el nivel de satisfacción entre los otros pacientes que ya experimentaron el efecto de un medicamento, abordando las Redes Neuronales para entender cómo la población general percibe la seguridad, reacciones y eficiencia de un medicamento. En el dominio particular de las farmacias, las reacciones positivas, negativas y neutrales son igualmente importantes para tomar una decisión sobre el uso de fármacos. Por otro lado, a partir de los resultados se demostró que el enfoque de creación de opinión basado en Redes Neuronales supera el método de la Máquina de Soporte Vectorial (SVM) en términos de precisión, exhaustividad y medida F. También, se encontró que los modelos de predicción basados en la Red Neuronal de Función de Base Radial (RBFN) se desempeñaron bien en todos los aspectos, mejor que la Red Neuronal Probabilística (PNN), a causa de las deficiencias que tienden a quedar atrapados en mínimos locales no deseados para alcanzar el mínimo global de un espacio de búsqueda muy complejo. De modo que, el enfoque empleado muestra un mejor resultado en términos de diversas medidas de rendimiento en comparación con otras revisiones de medicamentos de obras existentes recopiladas, además de ser una solución viable y óptima para aumentar el rendimiento de la clasificación.

Peng *et al.* [33] propusieron un método simple y eficaz para extraer ADE (*Adverse Drug Reactions*, Reacción Adversa a Medicamentos), mediante una canalización que consta de cinco pasos: 1) Captura de tweets; 2) Pre-procesamiento de datos; 3) Clasificación relacionada con medicamentos; 4) Análisis de sentimiento de tweets, y 5) Extracción de ADE. En este contexto, se identificó que las compañías farmacéuticas muchas de las veces no encuentran a tiempo los efectos secundaros que originan los fármacos, por el tamaño limitado de los ensayos clínicos, por tal motivo, las ADE no se identifican antes de salir al mercado, lo cual propicia que se retiren medicamentos en el mercado, lo que resulta en una pérdida financiera significativa para las compañías farmacéuticas, por lo tanto, es de gran importancia monitorizar y predecir las ADE. Sin embargo, los sitios Web de redes sociales, como Twitter, Facebook y Google+, proporcionan a los consumidores una forma de compartir experiencias con medicamentos que no reportan a los proveedores de atención médica ni a la FDA (*Food and Drug Administration*,

Administración de Alimentos y Medicamentos). Por otra parte, se efectuaron pruebas donde se encontró que los resultados son favorecedores, porque se extrajo con éxito 1,239 ADE mediante la red social de Twitter, entre ellos el 22% nuevos, para cinco medicamentos en cuatro meses, que equivale a cinco veces más en comparación con los resultados del método de referencia.

Sabra *et al*. [34] propusieron un *framework* de Extracción Semántica y Evaluación de Sentimiento de Factores de Riesgo (SESARF), el cual combina y mapea los conceptos relevantes, encuentra adjetivos y adverbios que reflejan el nivel de gravedad, así mismo, un método para el enriquecimiento semántico de los factores de riesgo de Tromboembolismo Venoso (VTE) para analizar las narrativas clínicas de los Registros de Salud Electrónicos (EHR) y predecir un diagnóstico VTE mediante el uso del clasificador de Máquina de Soporte Vectorial (SVM). El VTE es el tercer trastorno cardiovascular más común, estudios demuestran que la tasa de mortalidad asociada con la Embolia Pulmonar (EP) y la tasa de VTE recurrente es inaceptablemente alta, sin embargo, los esfuerzos preventivos comienzan con medidas de estilo de vida para reducir el riesgo de VTE. Por otra parte, el *framework* propuesto se evaluó mediante el enfoque de Procesamiento de Lenguaje Natural (PLN) para valorar la precisión y exhaustividad. La evaluación se efectuó mediante el uso de tecnologías de la Web Semántica y el aprendizaje automático para identificar los factores de riesgo que son esenciales para el diagnóstico y distinguir los niveles de gravedad utilizando el Sistema de Lenguaje Médico Unificado (UMLS) de MetaMap y LOD (*Linked Open Data*, Datos Abiertos Vinculados). También, se realizaron tres análisis sobre los datos: 1) Generales; 2) Basados en el género, y 3) Basados en la edad, donde se encontró que los resultados de los grupos de edad no contribuyeron a ninguna observación específica. Además, se encontró la existencia de una fuerte asociación entre la aparición de VTE y la combinación de los siguientes tres factores de riesgo: Diabetes, obesidad y tabaquismo. Finalmente se demostró que la predicción es factible y con una precisión sin presentar síntomas de 54.5% y una exhaustividad del 85.7%.

Denecke y Deng [35] desarrollaron un análisis que identifica y describe las facetas, así como el potencial que ofrece la minería de opiniones en el contexto de la medicina, también, efectuaron una evaluación cuantitativa con respecto al uso de las palabras y la distribución del sentimiento de un conjunto de datos obtenido de narrativas clínicas y redes sociales médicas derivadas de

seis fuentes diferentes. Dado que las observaciones personales y las actitudes de un médico influyen en la toma de decisiones clínicas, es crucial el análisis de los registros médicos para la obtención de una visión completa del estado de salud del paciente, además, la minería de opiniones ayuda a determinar el impacto de los documentos escritos. Por otro lado, se realizó un método de extracción para cuantificar las partes del habla, determinar la frecuencia de aparición y calcular coincidencias de términos con léxicos de sentimientos. Finalmente, se encontró que los análisis clínicos escritos por enfermería son más subjetivos que los de los informes de radiología, pero aún más objetivos que los datos de los medios sociales, así mismo se identificó que los léxicos usados tienen una cobertura diferente, pero se encuentran muchos más términos de sentimiento usando la herramienta SentiWordNet.

Wu *et al.* [36] propusieron un enfoque neuronal como representación jerárquica de tweets con el fin de reducir la influencia de errores ortográficos masivos y abreviaturas creadas por los usuarios de Twitter, por ello, el enfoque se centró en aprender las representaciones de las palabras de las personas y luego aprender las representaciones de los tweets. Así mismo, incluyó un modelo de auto atención con múltiples cabezas para detectar conjuntamente los tweets que mencionan nombres de medicamentos y las Reacciones Adversas a los Medicamentos (ADE). También, los autores utilizaron un mecanismo de atención aditiva para seleccionar las palabras informativas y así aprender más representaciones de tweets informativos. Se identificó que la red social de Twitter proporciona tweets que se recopilan fácilmente en tiempo real y la cantidad de tweets es enorme. Por lo tanto, detectar los tweets que mencionan los nombres de los medicamentos y los ADE ayuda a descubrir consecuencias graves o desconocidas del uso de medicamentos que no están cubiertos por los registros médicos. Para abordar la representación de las palabras se crearon tres sub-módulos, el primero es una capa de incrustación de caracteres, el segundo es una CNN (*Convolutional Neural Network*, Red Neuronal Convolucional) que proporciona pistas importantes para la detección de ADE y los nombres de medicamentos. Por otra parte, el tercero, es la incrustación de palabras, para mejorar las representaciones incorporando información semántica enriquecida extraída de una gran colección de tweets. En cuanto a la representación del tweet se dividió en tres módulos: 1) Una Bi-LSTM (*Bi-directional Long Short Term Memory*, Red Bidireccional de Memoria a Corto Plazo); 2) Red de auto-

atención de múltiples cabezas, y 3) Una red de atención aditiva. Los resultados experimentales en dos conjuntos de datos de referencia muestran que el enfoque propuesto mejoró de manera efectiva el rendimiento de la detección del nombre del medicamento y el ADE mencionados en los tweets y supera de manera consistente muchos métodos de referencia.

La farmacovigilancia es de gran importancia porque está relacionada con la detección, evaluación, comprensión y prevención de los Efectos Adversos a los Medicamentos (ADE) o con cualquier otro problema relacionado con los fármacos. En este contexto, Karimi *et al.* [37] presentaron un sistema, que lleva por nombre CSIRO Adverse Drug Event Miner (CADEminer), el cual extrae comentarios de consumidores en los foros sobre medicamentos, utilizando técnicas de búsqueda y Procesamiento de Lenguaje Natural (PLN) para extraer comentarios de efectos secundarios y otros conceptos relevantes como nombres de medicamentos y enfermedades. También, utilizó un enfoque basado en el aprendizaje automático que implementa campos aleatorios condicionales para la extracción de conceptos, así como el método de minería de reglas de asociación que se basa en el apoyo y la confianza de una regla potencial. De igual forma, se añadió un enfoque de recuperación de información para filtrar la información extraída que se conoce de un medicamento y así resaltar los efectos secundarios potencialmente desconocidos, mediante el uso de ontologías que ayudaron al mapeo entre los conceptos, tales como: SNOMED CT, AMT y Med-DRA. Finalmente, se identificó que CADEminer es útil para las agencias reguladoras, compañías farmacéuticas y para cualquier otra persona interesada en explorar la información sobre ADE.

Por otra parte, Khan *et al.* [38] realizaron una comparativa que revisa las diferentes técnicas de la minería de opiniones y destaca la necesidad de abordar los desafíos específicos del Procesamiento del Lenguaje Natural (PLN) para recopilar y examinar palabras relacionadas con la minería de opiniones. Así mismo, hicieron hincapié en la necesidad de conjuntos de datos estándar y metodologías de evaluación para mejorar los modelos que capturen el contexto y la proximidad. También, identificaron que las técnicas basadas en aprendizaje automático son de gran utilidad para la minería de opiniones y se implementa de dos formas; 1) Supervisadas, las cuales requieren datos etiquetados, y 2) Semi-supervisadas, que requieren un ajuste manual de expertos en el dominio. Con respecto la minería de opiniones se encontró que es un área de

campos de investigación diversificados que incluye aprendizaje automático, PLN, identificación del lenguaje y resumen de textos, donde los textos obtenidos de las revisiones se clasifican en diferentes idiomas, sin embargo, para cada lengua se identifican oraciones evaluativas y subjetivas. De modo que, al tener una relevancia cercana al PLN, la minería de opiniones enfrenta problemas como la resolución de co-referencia, el manejo de la negación y la desambiguación del sentido de las palabras.

Cabling *et al*. [39] llevaron a cabo un estudio sobre el impacto que conlleva implementar la minería de opiniones para los Grupo de Apoyo en Línea (OSG) sobre el cáncer de mama relacionado con el medicamento tamoxifeno, ya que es de gran importancia para comprender las emociones y las opiniones de los usuarios. El análisis se aplicó para 498 usuarios, donde los usuarios más activos equivalen a un 80% y el resto, a los usuarios menos activos. El objetivo de comparar los dos grupos es para la exploración de las posibles razones por las que los usuarios decidieron publicar un comentario y cómo su sentimiento juega un papel en la búsqueda o el suministro de soporte en línea en comparación con aquellos que no publicaron algún comentario. De modo que, cuanto más alta es la etapa de cáncer del usuario, menos probabilidades tiene de publicar y si decidiera publicar un comentario se centraría en los efectos secundarios y en la ansiedad o tristeza que lo acompaña, por otro lado, cuanto más baja es la etapa de cáncer del usuario, mayor es la probabilidad de publicar. Por tanto, se encontró que el análisis de los sentimientos de los usuarios proporciona una comprensión de cómo las interacciones específicas que promueven el apoyo llevan al desarrollo de dinámicas dentro del grupo y fuera de grupo, así como la comunicación hiper-personal dentro de los OSG. Sin embargo, a través de Big Data, la minería de opiniones y la investigación que explora el desarrollo de la cohesión grupal dentro de la Comunicación Mediada por Computadora (CMC), reveló una narrativa más rica sobre lo que ocurre en los OSG y proporciona una reflexión para que los moderadores estén atentos a la dinámica del sentimiento.

Actualmente la Web 2.0 permite a los individuos compartir datos valiosos y opiniones sobre productos o servicios comprados en línea, los datos compartidos y las opiniones no estructurados incluyen emociones, sentimientos, características, números, fechas y hechos, por lo cual, representa el centro de atención para la mayoría de los investigadores que intentan

recopilar y captar sentimientos populares. Es así que Solangi *et al*. [40] realizaron una revisión de las técnicas de Procesamiento del Lenguaje Natural (PLN) para la extracción de opiniones. Por otro lado, el estudio identificó las etapas del pre-procesamiento, requeridas para estructurar los textos, las cuales son: extracción de características, segmentación, tokenización, etiquetado gramatical y análisis en la extracción de opinión, donde la tokenización es una estrategia esencial para la mayoría de las tareas relacionadas con PLN, pero para los dialectos chinos, japoneses, entre otros, las palabras se componen de forma diferente. Por ello, se abordaron las herramientas que se ocupan para la segmentación de palabras chinas y la tokenización, tales como: 1) Fudan NLP en lenguaje JAVA; 2) La Plataforma de Tecnología del Lenguaje (LTP) en el sistema C ++ de código abierto para el análisis léxico; 3) Niu Parser en lenguaje C ++ que es un kit de herramientas para análisis semántico y sintáctico en chino; 4) Gensim Python, y 5) Stanford CoreNLP. Además, se encontró que la minería de opiniones tiene como objetivo extraer la orientación de sentimientos de los escritos, a través de la minería de opiniones se divide en tres niveles: nivel de documento, nivel de oración y nivel de grano fino. Finalmente, para el pre-procesamiento de texto se diseñaron técnicas que consisten en verificaciones para rastrear o clasificar los datos, sin embargo, la detección no siempre es la adecuada, por lo que, se sugirió realizar modificaciones a las técnicas de la PLN u otras técnicas relevantes, para que el usuario o beneficiario proporcione y reciba datos de información adecuados.

Luna-Aveiga *et al*. [41] propusieron un método basado en la minería de opiniones para detectar la reacción emocional de los pacientes con enfermedad asmática sobre factores de riesgo, actividades físicas, entre otros conceptos. La enfermedad del asma es un problema de salud global que afecta a todos los grupos de edad, que va del 1% al 21% en adultos y hasta el 20% en los niños, sin embargo, el número de pacientes con síntomas aumentó en casi un 30% en los últimos 20 años. Por otra parte, se identificó la existencia del crecimiento de información sobre salud y manejo de enfermedades en foros, blogs, microblogs y redes sociales, concretamente en Twitter, al ser una herramienta poderosa para difundir experiencias y fomentar conversaciones de autocontrol de enfermedades. De modo que el método propuesto aprovecha las tecnologías de la Web Semántica, específicamente de las ontologías, para representar el dominio del asma, así como de la herramienta SentiWordNet para determinar la polaridad de los conceptos del

asma contenidos en los mensajes de Twitter. El método se compone de cuatro elementos: 1) Módulo de normalización; 2) Módulo de anotación semántica; 3) La ontología para la autogestión del asma, y 4) Un módulo de identificación de polaridad del sentimiento. También, efectuaron evaluaciones al enfoque propuesto mediante los tweets recopilados, que dieron resultados alentadores con una precisión del 82.95%, una exhaustividad del 82.27% y una medida F del 82.36%. Los resultados de la evaluación mostraron la efectividad del método *n-gram around* usado para la identificación de la polaridad del sentimiento en el dominio de la salud, además se halló que el enfoque ayuda a crear conciencia sobre el asma, motivando así el comportamiento adicional de búsqueda de ayuda.

Rathi *et al.* [42] realizaron un análisis en que se discute a detalle sobre la minería de opinión, cómo el valor de la polaridad se relaciona con algo positivo o negativo y cómo abordar las reseñas en los blogs escritos en lengua romana. En este contexto, se encontró que hay muchas tareas diferentes para la minería de opinión, tales como: la extracción de opiniones, la extracción de sentimientos y el análisis emocional, sin embargo, el aspecto importante de la acción es recopilar la información de los comentarios que se encuentran en los blogs o en otros sitios y luego descubrir el comportamiento de esa información. Pero, existen muchas opiniones en la web y algunas opiniones contienen palabras dependientes del contexto las cuales presentan una clasificación de polaridad, por lo tanto, para catalogar los sentimientos correctamente, se elimina la ambigüedad de las oraciones con la ayuda de la herramienta llamada desambiguación de polaridad de palabras, que se refiere a la identificación computacional de la polaridad de una palabra en un contexto dado. Así mismo, el estudio identificó que los trabajos existentes se centran en el idioma inglés y no en el idioma romano, por lo que se encontró que hay un alcance futuro para crear una base de datos que ofrezca el valor de opinión de palabras romanas para compararlas con el idioma inglés, lo cual va a aumentar la calificación de rendimiento de los sitios web relacionados con cualquier producto que incluya reseñas en idioma romano.

Salas-Zárate *et al.* [18] desarrollaron un método basado en minería de opiniones que detecta eficazmente aspectos relacionados con la diabetes en los tweets, utilizando ontologías para describir semánticamente las relaciones entre conceptos del dominio específico, en el idioma inglés. Donde el sentimiento de los aspectos se calcula considerando las palabras alrededor del

aspecto que se obtienen a través de los métodos de “N-gram”. Por otra parte, el análisis identificó que la diabetes es una afección crónica que ocurre cuando el cuerpo no produce o usa suficiente insulina, por tal motivo es una de las mayores emergencias de salud a nivel mundial. Se estima que hay más de medio millón de niños menores de 14 años que viven con diabetes, 415 millones de adultos tienen diabetes y se estima que otros 318 millones de adultos tienen una tolerancia a la glucosa disminuida, lo que los pone en alto riesgo de desarrollar la enfermedad en el futuro. Es así que las redes sociales como Twitter son un excelente recurso para los pacientes, ya que se conectan con personas que tienen una condición similar y experiencias similares, sin embargo, la búsqueda de opiniones es una tarea difícil, porque una simple búsqueda en Twitter usando "diabetes" devuelve miles de tweets, por lo tanto, se necesitan sistemas de resumen de opinión que utilicen análisis de sentimiento o tecnologías de extracción de opinión. Por lo cual, el enfoque de clasificación de sentimiento propuesto se divide en tres componentes principales: el módulo de pre-procesamiento para limpiar y corregir el texto, el módulo de anotación semántica para la detección de aspectos y el módulo de clasificación de sentimiento que calcula la polaridad de cada aspecto encontrado en el léxico de SentiWordNet. Finalmente, el conjunto de experimentos dio resultados que mostraron que el método “N-gram around” obtuvo los mejores resultados con una precisión de 81.93%, una exhaustividad de 81.13% y una medida F del 81.24%. También, los experimentos mostraron que el léxico del sentimiento general no es suficiente para capturar los significados en los textos de la salud, además, el método propuesto requiere una ontología que modele el domino para identificar los aspectos en el dominio de la diabetes.

2.2 Análisis comparativo de los trabajos relacionados

A continuación, en la Tabla 2.1 se presenta una comparativa de los trabajos relacionados analizados anteriormente, con el fin de identificar las diferencias y similitudes en las contribuciones y en los resultados obtenidos y finalmente contrastar los datos con el proyecto planteado a desarrollado.

Tabla 2. 1 Análisis comparativo de los trabajos relacionados.

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
Lee *et al.* [29]	Falta de vigilancia en el mercado para identificar efectos secundarios no descubiertos previamente.	Twitter Search API, Sistema de Lenguaje Médico Unificado (UMLS), WordNet, ConText v2.	Sí	Modelos basados en CNN para la clasificación de ADE en tweets con un enfoque semi-supervisado, que superan a los modelos de clasificación supervisada un +9.9% medida de F1.
Alayba *et al.* [30]	Falta de análisis en tweets en el idioma árabe, que incluyen temas sobre servicios de salud.	Word2vec.	No	Un conjunto de datos en árabe aplicando el aprendizaje automático, y resultados con una precisión de

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
				aproximadamente entre el 85% y el 91%.
Cavalcanti y Prudêncio [31]	Se centró en la identificación de experiencias y opiniones de pacientes sobre tratamientos o medicamentos.	Stanford CoreNLP tools, API de corrección ortográfica de Google, MetaMap Java API, SIDER, MedTagger, NegEX7 Java, Weka.	No	La solución que se planteó logró resultados competitivos, además, se encontró la facilidad que tiene para adaptarse a otros idiomas, ya que no requiere de datos etiquetados.
Gopalakrishnan y Ramaswamy [32]	Pocos estudios que indagan sobre el impacto de las redes sociales en los pacientes con problemas de salud.	Rapid Miner, Weka, KNIME (o Konstanz Information Miner).	Sí	Se encontró que los modelos de predicción basados en RBFN se desempeñan bien en todos los aspectos dando un mejor resultado en términos

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
				de diversas medidas de rendimiento.
Peng *et al.* [33]	La falta de identificación oportuna de las ADE propicia que se retire medicamento del mercado, lo que resulta en una pérdi-da financiera para las compañías farmacéuticas.	Tweepy, API de Twitter, HiveQL, MetaMap, kits de herramientas de lenguaje natural y expresiones regulares de Python (NLTK) y Weka.	No	Un método para extraer los ADE, mediante una canalización que consta de cinco pasos, aplicado a la red social de Twitter, que extrajo con éxito un promedio de cinco veces el número total de ADE, en comparación con el método de referencia.
Sabra *et al.* [34]	El VTE es el tercer trastorno cardiovascular más común y la tasa recurrente es	MetaMap, Linked Open Data, TREC CDS, I2b2, Diseasome, DBpedia, Bio2RDF	No	Un *Framework* de SESARF para analizar las narrativas clínicas de EHR que predice un diagnóstico VTE con

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
	inaceptablemente alta, por lo cual se requieren de esfuerzos preventivos.	y API SentiWordNet.		una precisión sin síntomas de 54.5% y con exhaustividad del 85.7%.
Denecke y Deng [35]	Las observaciones personales y las actitudes de un médico influyen en la toma de decisiones clínicas, por lo cual es importante el análisis de los registros médicos.	SentiWordNet.	No	Los análisis clínicos escritos por enfermería son más subjetivos, y los léxicos usados tienen una cobertura diferente, pero se encuentran muchos más términos de sentimiento usando la herramienta SentiWordNet.
Wu *et al.* [36]	Detectar los tweets que mencionan nombres de	No se especifica.	No	Un enfoque neuronal como representación jerárquica de tweets

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
	medicamentos y los ADE ayuda a descubrir consecuencias graves o desconocidas del uso de medicamentos que no están cubiertos por los registros médicos.			con el fin de reducir la influencia de errores ortográficos masivos y abreviaturas creadas por el usuario de los tweets que mejora de manera efectiva el rendimiento de la detección y supera de manera consistente los métodos de referencia.
Karimi *et al*. [37]	Los efectos secundarios de los medicamentos son responsables de pérdidas humanas y financieras, por lo cual, la	PubMed como servidor web.	Sí	El sistema CADEminer basado en la web, que extrae comentarios de efectos secundarios y conceptos relevantes como nombres de

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
	farmacovigilacia es indispensable para la detección, evaluación, comprensión y prevención de los ADE.			medicamentos y enfermedades.
Khan *et al.* [38]	Existe la necesidad de abordar los desafíos abiertos específicos del procesamiento del PLN.	No se especifica.	No	Un análisis y comparativa que identificó que el análisis de sentimien-tos enfrenta problemas como la resolución de correferencia, el manejo de la negación y la desambiguación del sentido de las palabras.

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
Cabling *et al.* [39]	Es importante una conceptualización más clara del papel que juegan las plataformas en línea en la mediación del apoyo relacionado con la salud, específicamente del cáncer de mama.	CasualConc y BeautifulSoup.	No	Un estudio que encontró que los OSG resaltan y enfatizan un sentimiento especifico, por lo que ayudan a alentar a las conexiones entre aquellos pacientes que también publican.
Solangi *et al*. [40]	La Web 2.0 permite compartir datos y opiniones no estructurados que incluyen emociones, sentimientos, características, números, fechas y	Fudan NLP, LTP en C++, Niu Parser en C++, Gensim Python y Stanford CoreNLP.	No	Una revisión de las técnicas de PLN y minería de opiniones, en donde se sugirió realizar modifi-caciones a las técnicas de la PLN u otras técnicas relevantes.

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
	hechos, por ello es de gran importancia para los investigadores que intentan recopilar y captar sentimientos populares.			
Luna-Aveiga *et al*. [41]	La enfermedad del asma es un problema de salud global que afecta a todos los grupos de edad, en los últimos 20 años el número de pacientes aumentó aproximadamente un 30%.	API de Twitter, SentiWordNet, Ontología para el Autocontrol del Asma (OASM), Hunspell, *Framework* GATE, Babelfy.	Sí	Un método basado en minería de opiniones para detectar la reacción emocional de los pacientes con enfermedad asmática, con resultados de una precisión del 82.95%, una exhaustividad del 82.27% y una medida-F del 82.36%.

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
Rathi *et al*. [42]	El impacto de los medios de comunicación creó oportunidades extraordinarias para que los ciudadanos compartan públicamente sus opiniones, pero creó graves cuellos de botella cuando se trata de dar sentido a las opiniones.	No especifica.	No	Un análisis en el cual se discute a detalle sobre la minería de opinión, en donde se encontró como alcance futuro la creación de una base de datos que ofrezca el valor de la opinión en romano y en inglés para aumentar la calificación del rendimiento de los sitios web que incluyan reseñas en romano.
Salas-Zárate *et al*. [18]	Un método de minería de opiniones que detecta aspectos relacionados con la diabetes en los	Hunspell, Stanford CoreNLP para Java, Babelfy, SentiWord-Net, ontología DDO.	Sí	El conjunto de experimentos dio resultados que mostraron que el método "N-gram

Autor	Enfoque / Problemática	Tecnologías utilizadas	Arquitectura	Resultados
	tweets, utilizando ontologías, en el idioma inglés.			around" obtuvo los mejores resultados con una precisión de 81.93%, una exhaustividad de 81.13% y una medida F del 81.24%.

Por consiguiente, después de efectuar un análisis exhaustivo de los artículos abordados previamente, se encontró que los trabajos [29], [30], [33] y [36] se centran en la identificación de ADE mediante el análisis de comentarios de redes sociales. Los siguientes trabajos usaron comentarios de ADE para llevar a cabo minería de opiniones, pero específicamente para pacientes con enfermedades tales como: TDAH, SIDA, diabetes, obesidad o tabaquismo, haciendo uso de métodos de aprendizaje automático, pero analizando narrativas clínicas de pacientes, sin embargo, presenta una gran desventaja, ya que estas narrativas son subjetivas por el especialista que escribe el análisis clínico y no por el paciente.

Por otra parte, el trabajo [39] atiende las necesidades de un grupo en línea especializado para el cáncer de mama relacionado con el medicamento tamoxifeno, pero la desventaja que presenta es que el análisis efectuado solo es para un tipo de enfermedad y para un solo tipo de medicamento, no obstante, el grupo en línea solo está disponible en el idioma inglés.

Es por ello, que esta propuesta, a diferencia de los trabajos mencionados, propone un enfoque para el análisis de opiniones de medicamentos para enfermedades crónico-degenerativas en idioma español, el cual, según la clasificación de Internet *World State*, es el tercer idioma más utilizado en Internet, dato que enfatiza la importancia del procesamiento de opiniones en este idioma.

2.3 Propuesta de solución

Con el objetivo de dar una solución adecuada al problema planteado, se efectuó un análisis de las herramientas para Web *Scraping*, extracción de subtítulos, minería de opiniones, editores de ontologías, sistemas gestores de base de datos, lenguajes de programación, entornos de desarrollo y metodologías y se eligieron las más idóneas para ayudar a la solución del problema planteado, se eligieron las herramientas que brindan ventajas de comunicación, funcionalidad y estabilidad para el desarrollo e integración.

En la Tabla 2.2 se muestra las herramientas que componen la propuesta de solución.

Tabla 2. 2 Propuesta de solución

Aspecto	Recurso
API Web Scraping	JSoup
API extracción de *captions*	JDOM
API de minería de opiniones	Stanford Core NLP
Editor de Ontologías	Protégé
***Framework* de ontologías**	Apache Jena
SGBD	PostgreSQL
Lenguaje de programación	Java
IDE	NetBeans
Metodología	UWE

A continuación, se describen las tecnologías que se emplean para esta tesis.

2.3.1 JSoup

Es una API para Java que trabaja con HTML, el cual resulta conveniente para extraer y manipular datos, utilizando el DOM (*Document Object Model*, Modelo de Objetos del Documento).

Se diseñó para tratar con todas las variedades de HTML que se encuentran en la Web, además, extrae datos de selectores de CSS (*Cascading Style Sheets*, Hojas de Estilo en Cascada), y provee un HTML ordenado, como salida [43].

2.3.2 JDOM ™

Es una biblioteca de código abierto para la manipulación de XML, está centrado en Java y optimizado para Java. Es una API completamente natural para los desarrolladores actuales de Java.

Si bien JDOM interactúa con los estándares existentes, como la simple API para XML (SAX) y el Modelo de objeto de documento (DOM), no es una capa de abstracción o mejora para esas

API's. Más bien, proporciona un medio robusto y liviano para leer y escribir datos XML sin las opciones complejas y de consumo de memoria que ofrecen las API's actuales [44].

2.3.3 Stanford Core NLP

Stanford CoreNLP es una API que provee un conjunto de herramientas de tecnología de lenguaje humano. Permite indicar las formas básicas de las palabras, sus partes del habla, ya sean nombres de compañías, personas, entre otros, así mismo, normaliza los datos no estructurados, lo cual, se refiere a fechas, horas y cantidades numéricas.

Esta API marca la estructura de las oraciones en términos de frases y dependencias sintácticas, indica sentimientos y extrae relaciones particulares o de clase abierta entre las menciones a las entidades.

En la Figura 2.1 se muestra un ejemplo del proceso que aplica la API en una oración para identificar una entidad.

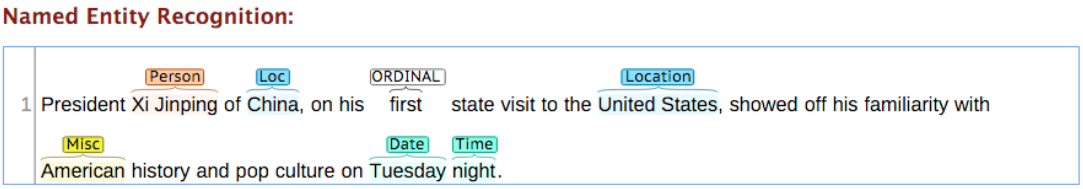

Figura 2. 1 Ejemplo de reconocimiento de una entidad nombrada

La API de Stanford CoreNLP es útil cuando se requiere de:

- Un conjunto de herramientas de PLN integrado con una amplia gama de herramientas de análisis gramatical.
- Un anotador rápido y robusto para textos arbitrarios, ampliamente utilizado en la producción.
- Un paquete moderno, actualizado regularmente, con el análisis de texto de más alta calidad en general.
- Soporte para varios idiomas.

- APIs disponibles para la mayoría de los principales lenguajes de programación modernos.
- Capacidad para funcionar como un simple servicio web [45].

2.3.4 Protégé

Es un editor y marco de ontologías de código abierto y gratuito para construir sistemas inteligentes, el cual, cuenta con el respaldo de una sólida comunidad de usuarios académicos, gubernamentales y corporativos, que utilizan Protégé para crear soluciones basadas en el conocimiento en áreas tan diversas como la biomedicina, el comercio electrónico y el modelado organizativo.

Proporciona a los desarrolladores la facilidad de integrar la salida de Protégé con sistemas de reglas u otros solucionadores de problemas para construir una amplia gama de sistemas inteligentes. Así mismo, Protégé es totalmente compatible con las últimas especificaciones OWL 2 y RDF del consorcio de la World Wide Web (W3C) [46].

2.3.5 Apache Jena

Apache Jena es un marco Java gratuito y de código abierto para crear aplicaciones de datos semánticos y Web Semántica. El marco está compuesto por diferentes API que interactúan entre sí para procesar datos RDF.

También, provee herramientas de líneas de comando para realizar tareas clave con Jena, sin embargo, Jena es usado en programas escritos en Java [47].

2.3.6 PostgreSQL

PostgreSQL es un potente sistema gestor de base de datos relacional de objetos de código abierto que tiene una sólida reputación de confiabilidad, solidez de funciones y rendimiento. Cuenta con una gran cantidad de información que describe cómo instalar y usar PostgreSQL a través de la documentación oficial. Además, proporciona características destinadas a ayudar a los desarrolladores a crear aplicaciones, a los administradores a proteger la integridad de los datos y a crear entornos tolerantes a fallas, así como, a administrar los datos sin importar qué tan grande o pequeño sea el conjunto de datos. PostgreSQL es gratuito y de código abierto, también

es altamente extensible, es decir, permite definir sus propios tipos de datos y construir funciones personalizadas [48].

2.3.7 Java

Java es considerado según la empresa SUN como un lenguaje sencillo, orientado a objetos, distribuido, interpretado, robusto, seguro, independiente de las arquitecturas, portable, eficaz, multihilo y dinámico. Donde, un programa desarrollado en Java es interpretado por una máquina virtual JVM (*Java Virtual Machine*, Máquina Virtual de Java), lo cual conlleva al hecho de no tener que recompilar el programa en cada sistema operativo. Además, es un lenguaje fuertemente tipado y estricto, el cual verifica la sintaxis del código en el momento de la compilación, así como también en la ejecución. Java es compatible con sistemas operativos Windows, Linux y Mac OS.

El diseño de Java aporta:

- Lenguaje de sintaxis sencilla, orientada a objetos que permite optimizar el tiempo de desarrollo, compilación y ejecución.
- Las aplicaciones son resistentes porque el motor de ejecución se encarga de la gestión de la memoria.
- Las aplicaciones gráficas son eficaces debido a la puesta en marcha y a la asunción del funcionamiento de varios procesos, mejor conocido como Thread y multithreading.
- Java se encarga de que el Applet de las aplicaciones no realice ninguna manipulación u operación peligrosa [49].

2.3.8 NetBeans

El IDE (*Integrated Development Environment*, Entorno Desarrollo Integrado) de NetBeans es reconocido como un IDE oficial para el lenguaje de programación Java, el cual, con sus editores, analizadores de código, así como sus convertidores actualiza las aplicaciones de manera rápida y sin problemas para usar nuevas construcciones de lenguaje Java 8, como lambdas, operaciones funcionales y referencias de métodos.

Así mismo, el IDE proporciona analizadores de lotes y convertidores para buscar a través de múltiples aplicaciones al mismo tiempo. Por otra parte, permite re factorizar fácilmente el

código, con una gama de herramientas útiles y poderosas, mientras que también provee de plantillas de código, sugerencias de codificación y generadores de código [50].

2.3.9 UWE

UWE (*Unified Modeling Language Web Engineering*, Ingeniería Web del Lenguaje Unificado de Modelado) es una metodología que permite especificar de mejor manera una aplicación Web en su proceso de creación. Mantiene una notación estándar basada en el uso de UML (*Unified Modeling Language*, Lenguaje Unificado de Modelado) para sus modelos y sus métodos, lo que facilita la transición. La metodología define claramente la construcción de cada uno de los elementos del modelo.

En su implementación es necesario contemplar las siguientes etapas y modelos:

1. Análisis de requisitos: Plasma los requisitos funcionales de la aplicación Web mediante un modelo de casos de uso.
1. Modelo de contenido: Define, mediante un diagrama de clases, los conceptos a detalle involucrados en la aplicación.
2. Modelo de navegación: Representa la navegación de los objetos dentro de la aplicación y un conjunto de estructuras como son índices, menús y consultas.
3. Modelo de presentación: Representa las interfaces de usuario por medio de vistas abstractas.
4. Modelo de proceso: Representa el aspecto que tienen las actividades que se conectan con cada proceso.

Cada uno de estos modelos se representa como paquetes UML, dichos paquetes son procesos relacionados que facilitan su refinamiento a través de iteraciones sucesivas durante el desarrollo del UWE como se muestra en la Figura 2.2 [51].

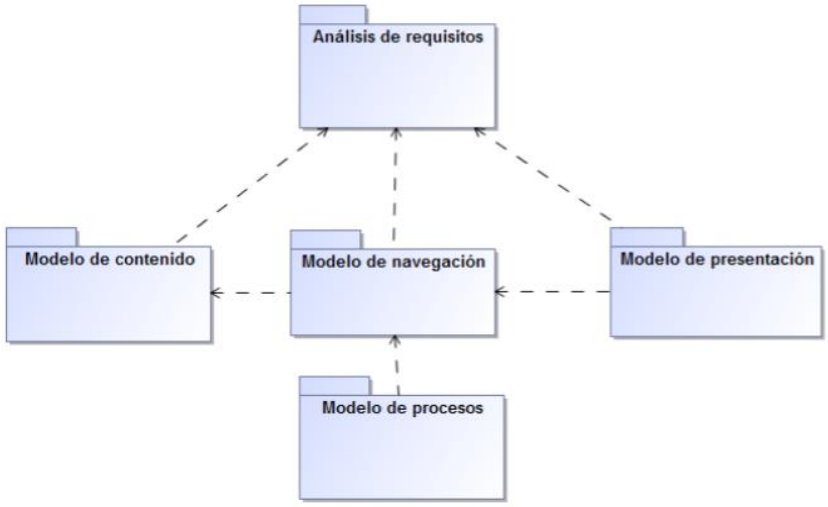

Figura 2. 2 Modelo de la metodología UWE

Se optó por la alternativa de solución antes mostrada, ya que se considera como la más adecuada, y aquella que ofrece más ventajas. Es así, que se eligieron las API's JSoup, JDOM y Stanford Core NLP por ser de las más completas, estables y con una amplia comunidad de desarrollo, por ello, tienen una extensa documentación. La elección de las API's conllevó a seleccionar como lenguaje de desarrollo Java, por ser un lenguaje con gran potencial para el desarrollo web por sus marcos de trabajo y además por ser gratuitos. Además, se eligieron los *frameworks*, Apache Jena y Protégé, por ser herramientas muy potentes y estables que cuentan con una comunidad sólida de usuarios académicos, gubernamentales y corporativos.

Por otra parte, se optó por usar NetBeans dado el soporte y la documentación vasta que provee, además por ser un entorno intuitivo y potente para desarrollar diversos tipos de aplicaciones. PostgreSQL es idóneo porque es robusto y es ampliamente utilizado con el lenguaje Java, además necesita muchos requerimientos del sistema operativo donde se emplee, por lo cual es un SGBD muy recomendado para usar junto con Java. Finalmente, la metodología UWE es adecuada para emplear en el proyecto de desarrollo, ya que es destacada por ser orientada a objetos, además de tener definidas claramente sus fases y sus entregables, también, por seguir bajo la notación UML, lo cual representa un estándar que da un valor agregado al desarrollo del proyecto propuesto.

Capítulo 3. Aplicación de la metodología

A continuación, en el presente capítulo de describe el proceso de desarrollo de la aplicación Web que da solución al problema planteado en el capítulo 1 dicho desarrollo se lleva a cabo a través de las fases que proporciona la metodología UWE.

3.1 Metodología de desarrollo

3.1.1 Análisis de requisitos

El análisis de los requisitos permite entender cómo el usuario interactúa con el sistema y con ello dar una solución adecuada al planteamiento de problema. Por ello, es importante conocer al usuario que usará la aplicación Web y el uso que le dará.

A continuación de describen las funciones que realiza en la aplicación Web, es decir, el cómo interactúa el usuario con las funciones, descritos mediante casos de usos, como se muestra en la Figura 3.1.

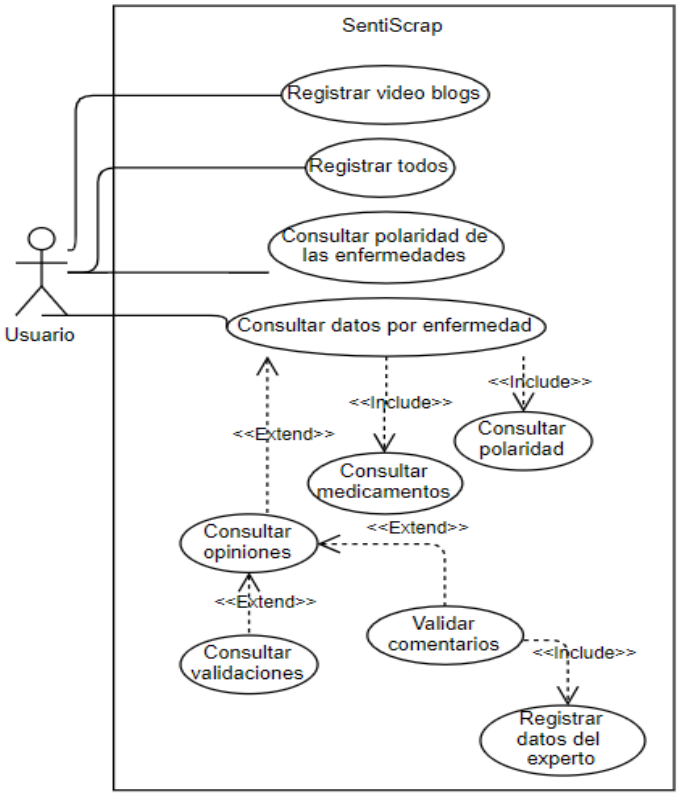

Figura 3. 1 Diagrama de casos de uso de la aplicación Web

El usuario que acceda a la aplicación Web es capaz de registrar un video blog, así como un foro o blog especializado que quiera analizar. También, consultar la información que contienen medicamentos prescritos para cada enfermedad, disponibles en la aplicación Web.

Por otra parte, cuando consulta los medicamentos el usuario también tiene la disponibilidad de consultar la polaridad de cada medicamento, los efectos adversos a los medicamentos encontrados y consultar los comentarios.

Cuando el usuario consulte los comentarios también visualizará las validaciones realizadas por los expertos en el dominio, incluso si el usuario es un experto tiene la opción de validar el comentario, con un registro previo de sus datos.

3.1.2 Modelo de contenido

En este apartado se presenta una especificación visual de la composición de la aplicación Web, representado mediante un diagrama de clases, como se muestra en la Figura 3.2 y Figura 3.3.

El paquete que se encarga de la recolección de los datos contiene las clases: Foros, VideoCaptions y CaptionsModelo para realizar el Web *Scraping* en los foros, sitios Web especializados y video blogs. El paquete sentiscrap.preprocesamiento contiene una clase para la normalización y limpieza del texto recolectado. Por otra parte, el paquete sentiscrap.dominio cuenta con la clase BoW que permite acceder a la ontología. El paquete de procesamiento contiene la clase para implementar el análisis de polaridad a través de la clase EntrenaAlgoritmo. También el paquete sentiscrap.valexpert.ws contiene una clase que se implementa para la validación de los comentarios que realizan los expertos en el dominio. Finalmente, los paquetes sentiscrap.aplic.dtos, sentiscrap.aplic.modelo, sentiscrap.aplic.daos, sentiscrap.aplic.repositorio son el modelo base de la aplicación que funcionan como medio para la conexión hacia la base de datos.

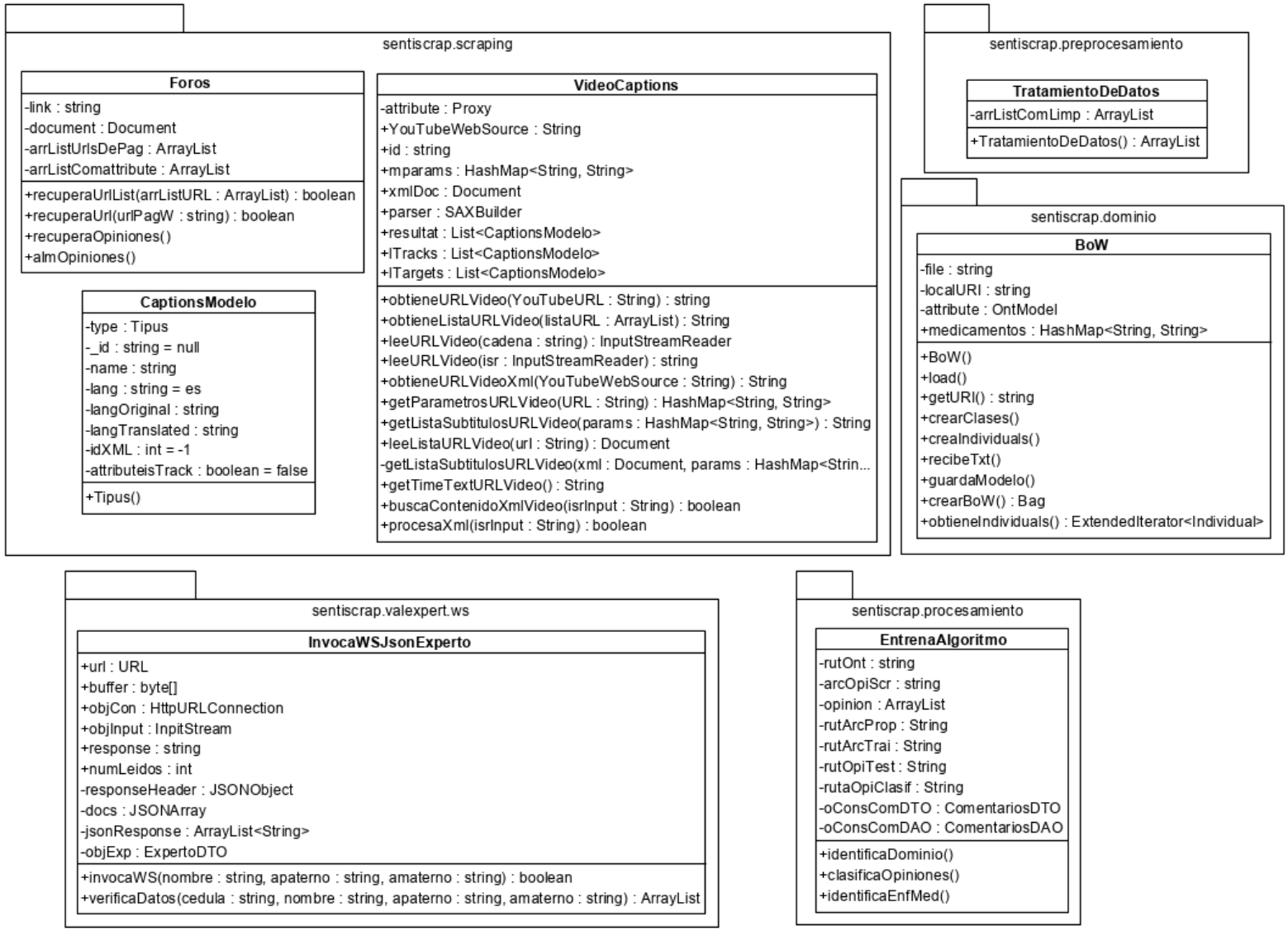

Figura 3. 2 Diagrama de clases de la aplicación Web parte 1

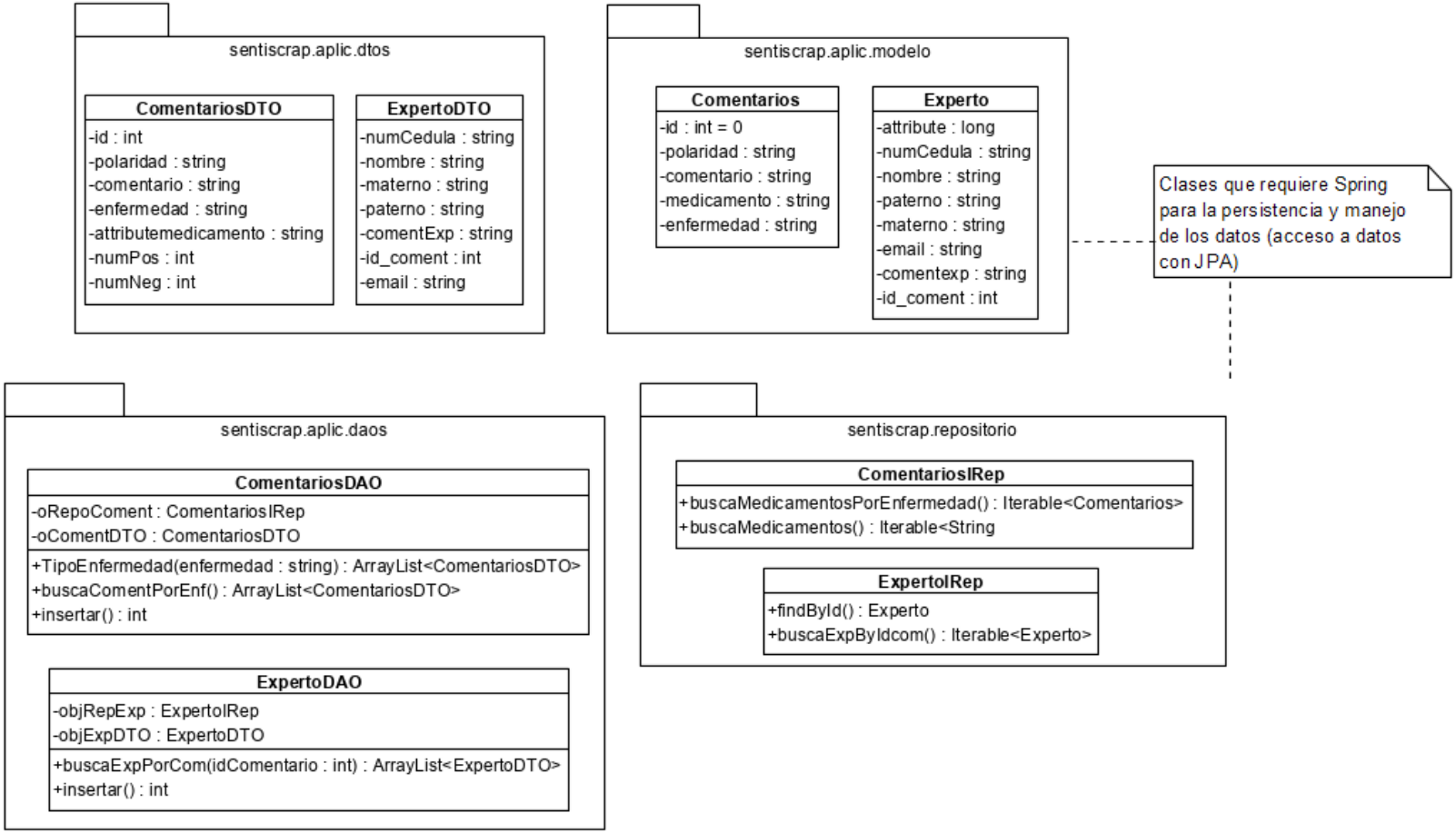

Figura 3. 3 Diagrama de clases de la aplicación Web parte 2

3.1.3 Modelo de navegación

Este modelo permite establecer un mapa de navegación para la aplicación Web, a través de la representación de un diagrama de clases estereotipado, que permite la visualización de las relaciones existentes entre todas las clases y el flujo permitido en el que el usuario navegará en la aplicación.

El mapa de navegación inicia desde la página de inicio, con un menú con cuatro posibles opciones para iniciar la navegación: Hepatitis, Diabetes, Hipertensión, AñadirFuentes y Ver Estadística. Si el usuario decide iniciar con cualquiera de las tres opciones de enfermedades, la aplicación lo llevará a la clase en la cual se presentan los medicamentos referidos para cada enfermedad y de manera global se mostrará la polaridad. Si el usuario lo requiere, continuará para ver más detalles y se le mostrará un botón para acceder a los comentarios por enfermedad. Al acceder a dichos comentarios, el usuario tendrá la opción de elegir ver las validaciones o validar un comentario, el cual lo dirige a la "RegistrarDatos".

Por otra parte, si el usuario en la página de inicio, comienza la navegación con la opción "AñadirFuentes", se desplegará un menú con dos opciones "Foros" o "Videoblogs" que dirigen al usuario a una clase donde procesa los URL a analizar. Finalmente, el usuario tiene la opción de navegar mediante la opción de "Ver Estadistica" la cual muestra los datos de polaridad de los tres tipos de enfermedad abordados. A continuación, en la Figura 3.4 se muestra el mapa de navegación.

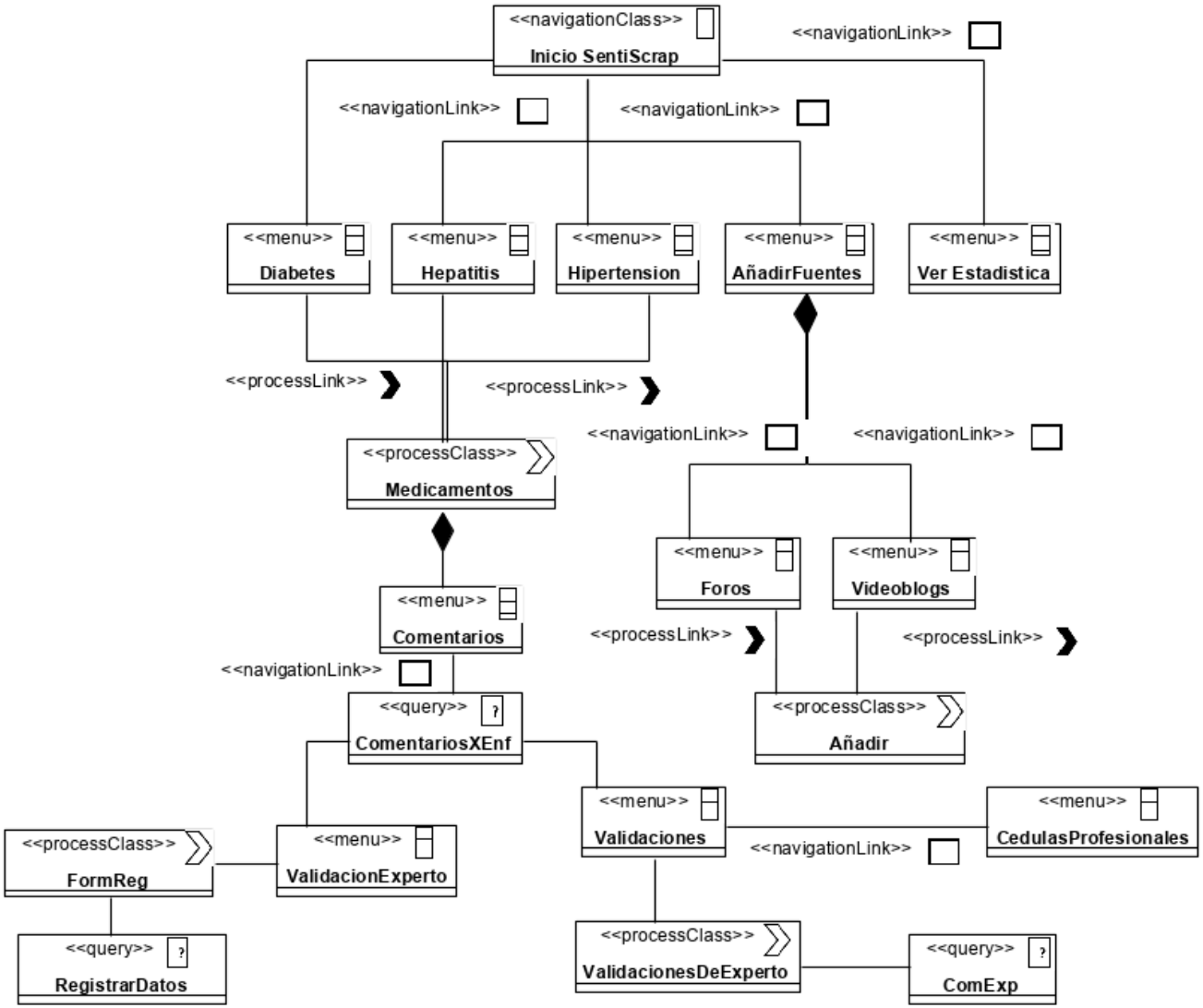

Figura 3. 4 Diagrama de navegación de la aplicación Web

3.1.4 Modelo de presentación

El objetivo principal de este modelo es identificar los elementos de la aplicación Web, así como ofrecer una visión clara de la interfaz de usuario. A continuación, se mostrará el maquetado de la aplicación a través de *Mockups* realizados a partir del diagrama de navegación, mostrado en la sección previa. Donde cada *Mockup* representa una clase con elementos tales como: texto, botones, imágenes, formularios, entre otros.

En la Figura 3.5 se muestra la página principal con un mensaje de inicio, así como un menú con cinco opciones y un pie de la página con texto que indica fecha y lugar de realización.

Figura 3. 5 *Mockup* de página inicio

A continuación, la Figura 3.6 representa el maquetado que contiene un carrusel de imágenes de medicamentos para la enfermedad de “Hepatitis”, la polaridad del medicamento representada por emoticones, además, un botón para ver más detalles.

Figura 3. 6 *Mockup* para la opción “Hepatitis”

La Figura 3.7 muestra una gráfica que representa la polaridad positiva y negativa respecto al medicamento seleccionado, así como un título y un texto descriptivo. Además, se sigue conservando el menú con cinco opciones y el pie de la página.

Figura 3. 7 *Mockup* de polaridad por medicamento

El siguiente maquetado corresponde a la representación de los comentarios descritos en texto sobre cada enfermedad y medicamento. Además, se muestra un elemento menú con dos opciones que dirigen hacia dos *Mockups* más. Por otra parte, se sigue conservando el menú que dirige a solo dos únicas opciones, como se muestra en la Figura 3.8.

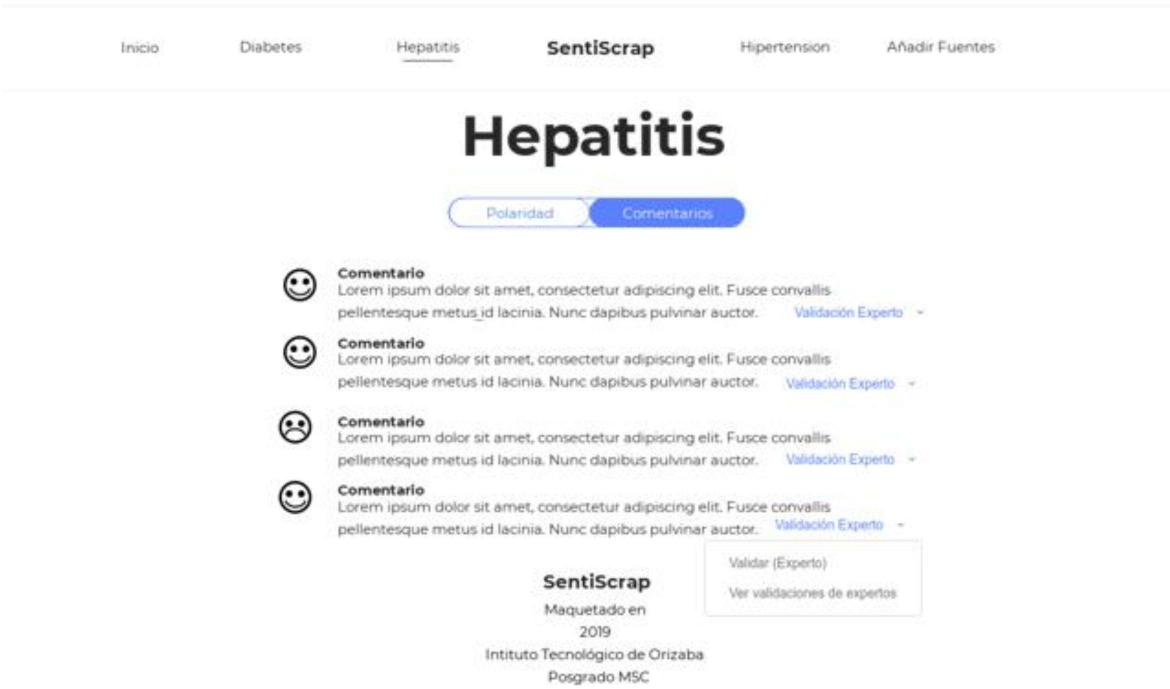

Figura 3. 8 *Mockup* de comentarios por medicamento

En la Figura 3.9 corresponde al *Mockup* con un formulario para el registro del nombre completo del experto y de la cédula, así como un botón para la efectuar la acción de “Enviar”.

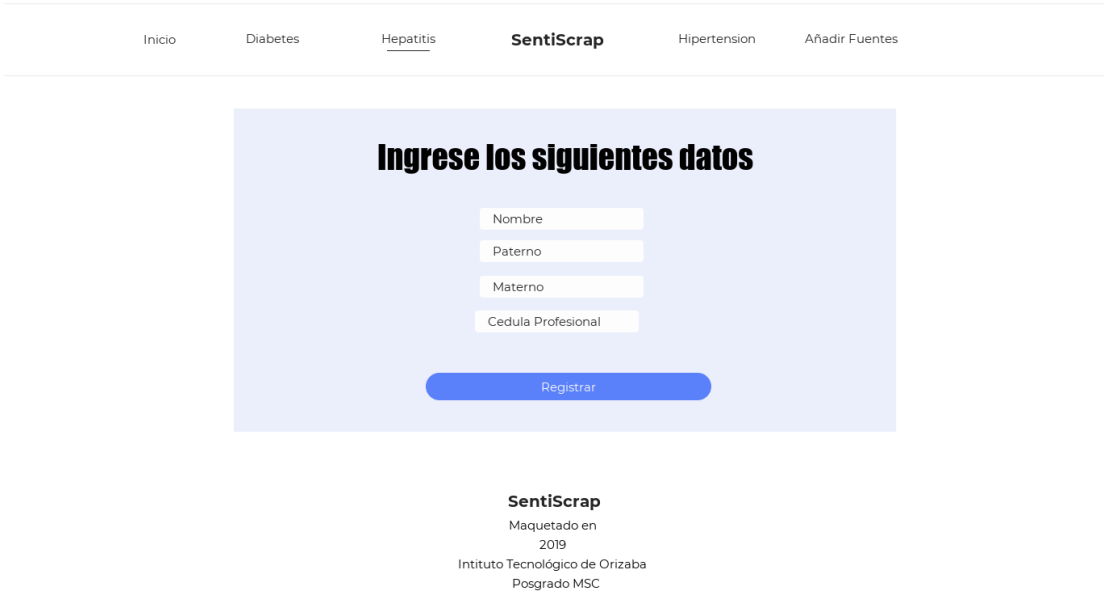

Figura 3. 9 *Mockup* de registro del experto

El maquetado de la Figura 3.10 contiene un elemento de tipo texto donde el usuario ingresa su comentario, así como un botón para la efectuar la acción de “Enviar”.

Figura 3. 10 *Mockup* de registro de comentario del experto

A continuación, en la Figura 3.11, se muestra el *Mockup* para especificar los comentarios realizados por el usuario y un *link* que dirige hacia un sitio Web externo.

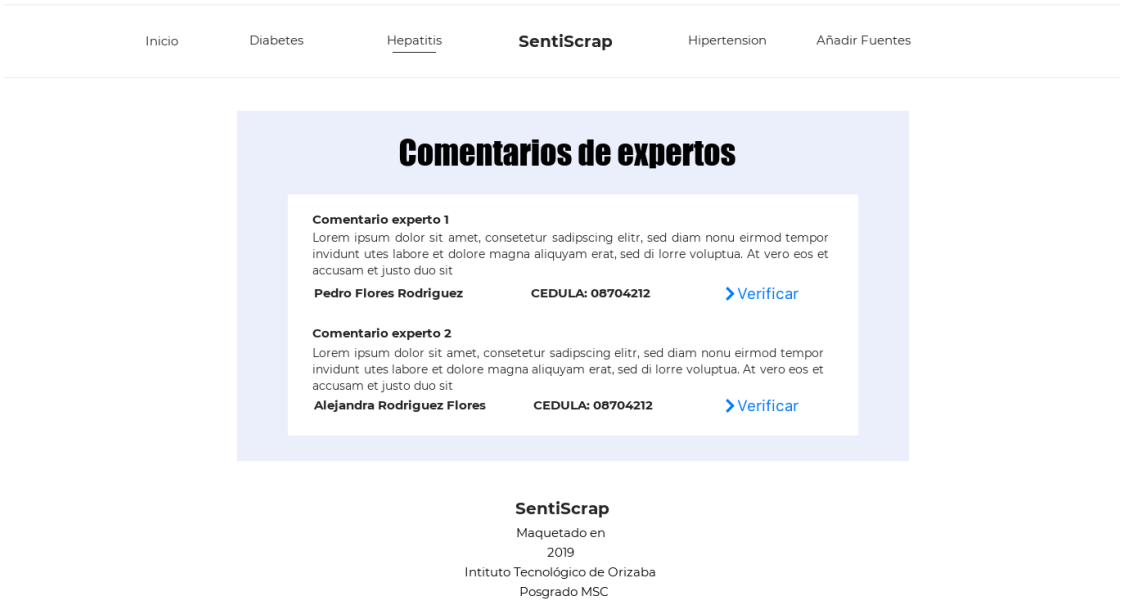

Figura 3. 11 *Mockup* de comentario del experto

La Figura 3.12 corresponde al *Mockup* "Añadir Fuentes" el cual presenta como componentes un menú con dos opciones para "Foros" y "Videos", un formulario con un campo para añadir una URL correspondiente a foros y un botón para efectuar la acción de "Añadir".

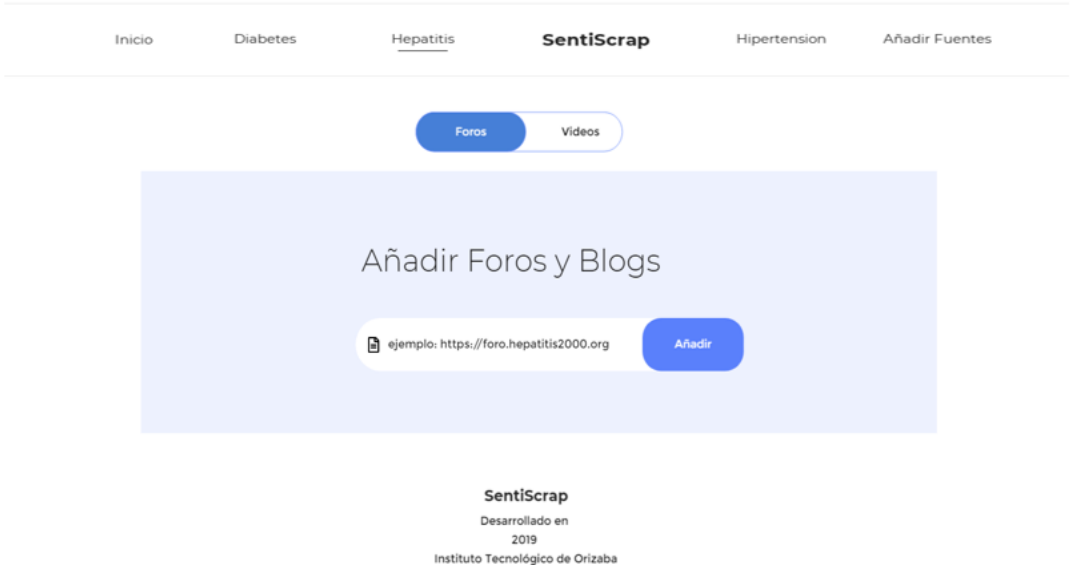

Figura 3. 12 *Mockup* para añadir fuentes de foros

El *Mockup* de la Figura 3.13 “Añadir Fuentes” continúa presentando el componente menú con dos opciones, así mismo, un formulario con un campo para añadir una URL correspondiente a los videoblogs y un botón para efectuar la acción de “Añadir”.

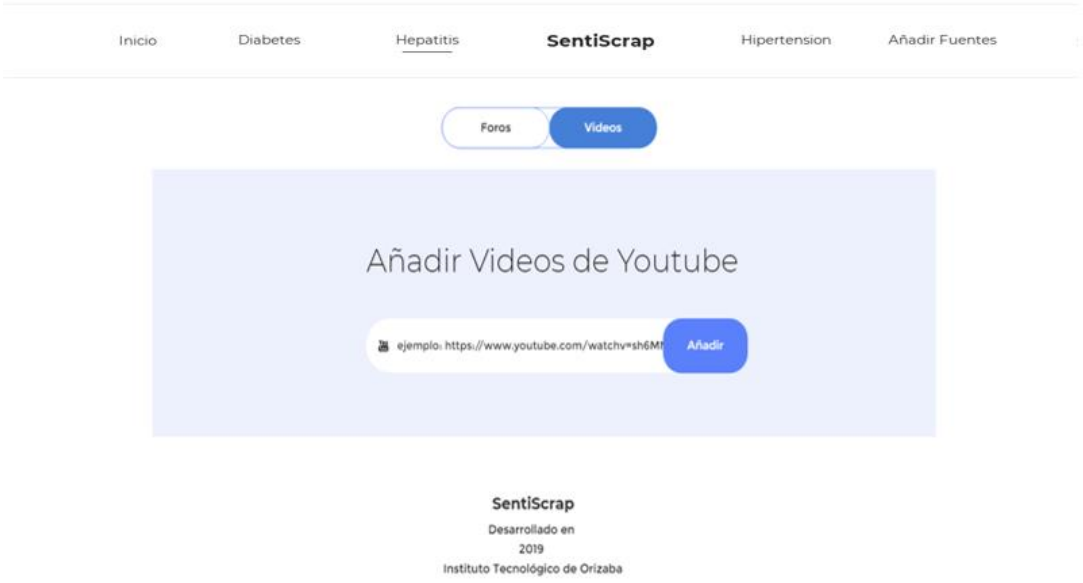

Figura 3. 13 *Mockup* para añadir fuentes de videoblogs

3.1.5 Modelo de procesos

Este modelo describe de forma más específica los datos de entrada y salida en los distintos escenarios que presenta la aplicación Web. Es decir, representa el aspecto que tienen las acciones de las clases del proceso. Para representar dichos procesos, se incluyen los siguientes diagramas de actividades estereotipados de acuerdo a las especificaciones de la metodología empleada.

En la Figura 3.14 se muestra el proceso de la aplicación a través de las actividades para el caso en el cual el usuario interactúa en la aplicación con la opción para la enfermedad “Hepatitis”, sin embargo, este mismo proceso se repite para las demás enfermedades incluidas en la aplicación Web.

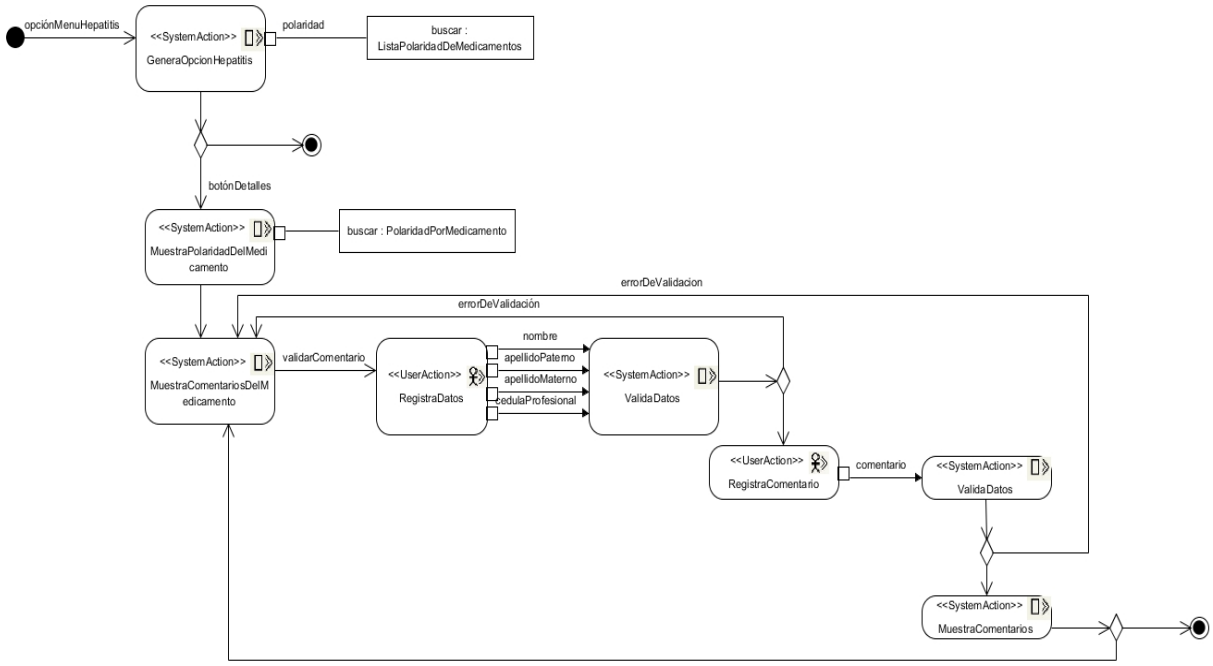

Figura 3. 14 Diagrama de actividades de la opción “Hepatitis”

De igual forma, en la Figura 3.15 se muestra el proceso de la aplicación a través de las actividades para el caso en el cual el usuario interactúa en la aplicación para añadir nuevas fuentes para procesar y analizar. Se presenta el caso para añadir URL’s de videoblogs el cual es igual para el caso de foros y sitios Web especializados.

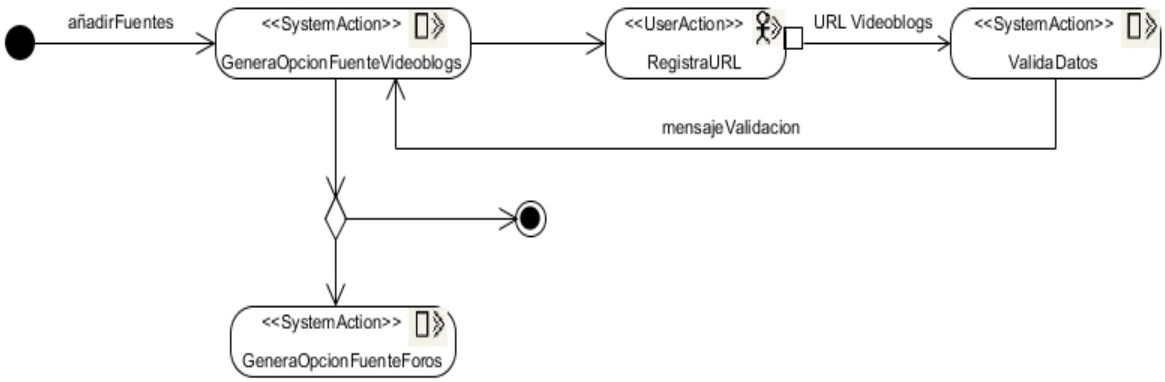

Figura 3. 15 Diagrama de actividades para añadir fuentes de tipo Videoblogs

3.2 Arquitectura

La arquitectura de un sistema permite identificar de forma clara y precisa la solución que se aplica a un problema planteado. Por ello, se realizó una arquitectura que cuenta con siete módulos con sus respectivas actividades, que dan solución óptima al problema antes planteado, como se muestra en la Figura 3.16.

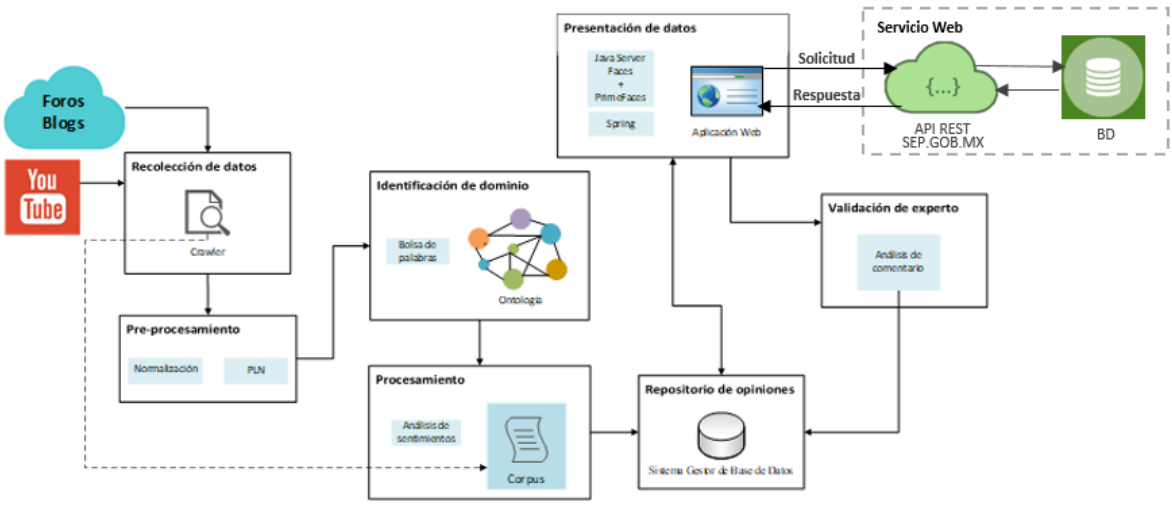

Figura 3. 16 Arquitectura de la aplicación Web

3.2.1 Módulo de recolección de datos:

En este módulo se realiza la búsqueda en la Web, específicamente en foros, blogs y video blogs, con el propósito de extraer los comentarios que mencionen enfermedades crónico-degenerativas, medicamentos y síntomas referidos. En la Figura 3.17 se muestra un diagrama representativo al flujo del módulo presente.

Figura 3. 17 Diagrama de flujo del Web Scraping

3.2.2 Módulo de pre-procesamiento:

El pre-procesamiento de los datos es un paso importante para la normalización del texto, por lo tanto, se optó por utilizar tres fases para el tratamiento de los datos.

1. Borrar caracteres inusuales: los comentarios contienen caracteres especiales que no proporcionan información, por lo que se borran.
2. Eliminar comentarios duplicados: Este paso es importante, ya que los comentarios duplicados afectan al resultado final del análisis, por lo que es importante asegurarse de que se eliminen los comentarios duplicados.
3. Borrar comentarios que sólo tienen URLs: los comentarios que sólo incluyen enlaces a otros sitios no contribuyen como comentario para analizar la polaridad, por esta razón son descartados.

La aplicación de estas tareas a los comentarios asegura un mejor análisis de los sentimientos, sin embargo, el uso incorrecto del lenguaje es un escenario común, creado por el uso de abreviaturas o errores ortográficos por parte de los usuarios, requiriendo un mayor esfuerzo para llevar a cabo actividades de minería de opinión.

3.2.3 Módulo de identificación de dominio:

La identificación de palabras tiene alta relevancia porque permite verificar que los comentarios obtenidos mencionen a los medicamentos prescritos para las enfermedades crónico-degenerativas, lo que proporciona como resultado un conjunto de datos más específico, con información valiosa para el análisis.

3.2.4 Módulo de procesamiento:

Este módulo adopta el enfoque de aprendizaje automático supervisado que utiliza un corpus etiquetado semiautomático, necesario para entrenar el algoritmo que permite el análisis de sentimientos, lo cual, ayuda a reconocer nuevas opiniones en el idioma español y clasificarlas correctamente.

3.2.5 Módulo de repositorio de opiniones:

Las opiniones y la polaridad de los medicamentos resultantes del análisis se almacenan en una base de datos para mantener los datos disponibles para su posterior consulta.

3.2.6 Módulo de validación de experto:

Se realiza un análisis de los comentarios realizados por los expertos de la salud que validan los comentarios recopilados de foros, blogs y video blogs, para así determinar los comentarios que se encuentran verificados de forma positiva y negativa.

En la Figura 3.18 se muestra de forma conceptual en qué consiste la validación del experto, es decir, el especialista en el dominio valida los comentarios mostrados obtenidos del Web Scraping, pero para ello es necesario que proporcione su nombre y cédula profesional que se valida a través de un servicio Web y posteriormente se le habilitará la opción de escribir un comentario.

Figura 3. 18 Diagrama de flujo de la validación y comentario del experto

3.2.7 Módulo de presentación de datos:

Se presenta una interfaz Web que apoya la interacción de los usuarios con la aplicación Web con el objetivo de conocer los comentarios y la polaridad de los fármacos prescritos para las enfermedades crónico-degenerativas, concretamente la diabetes mellitus, la hepatitis y la hipertensión.

3.3 Diseño de la base de datos

El diseño de la base de datos a través del diagrama entidad – relación muestra la especificación de los datos que se emplean en la aplicación Web, así como de la relación que tienen unos datos con otros, respecto a las entidades definidas, como se muestra en la Figura 3.19.

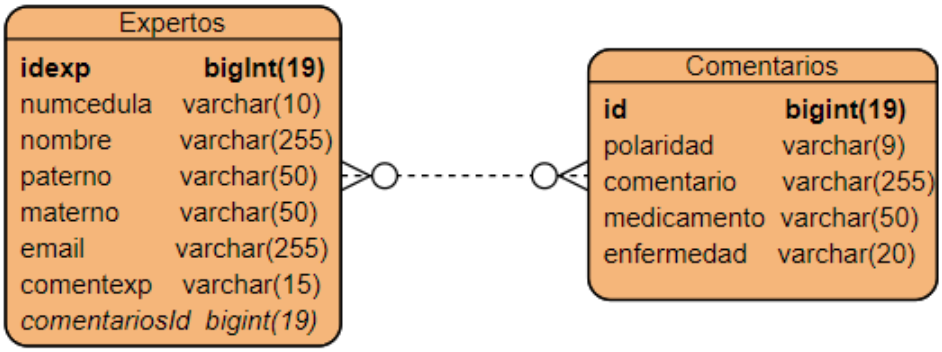

Figura 3. 19 Diagrama E-R de la base de datos para la aplicación Web

3.4 Metodología de la ontología

3.4.1 Proceso de creación de ontologías Methontology

Para el diseño de la ontología se llevaron a cabo una serie de actividades de conceptualización las cuales se dividen en 11 tareas, como se muestra en la Figura 3.20. Cada tarea se centra en un aspecto del modelo conceptual del conocimiento: términos, taxonomía, relaciones, axiomas, reglas y dependiendo del autor, algunos elementos tienen una aproximación matemática.

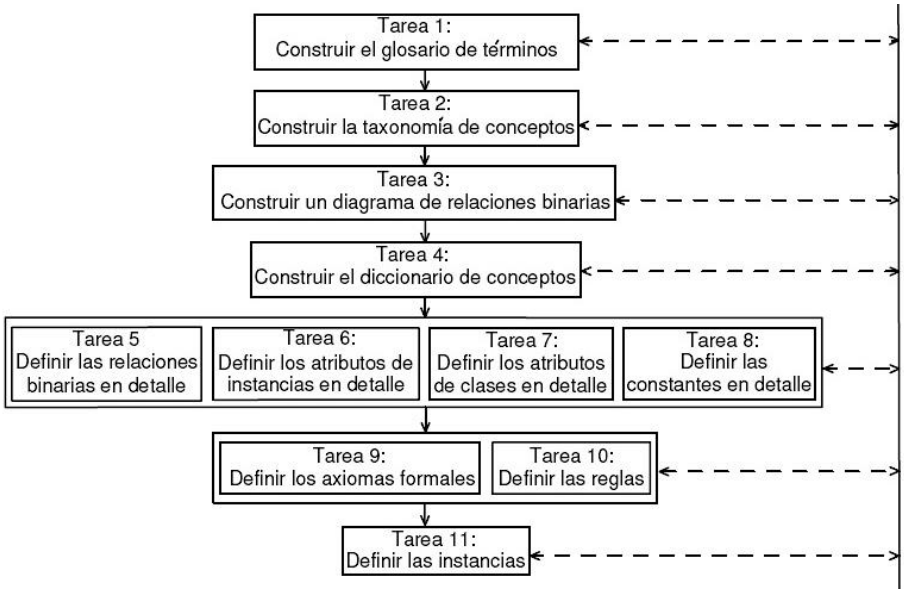

Figura 3. 20 Lista de actividades de conceptualización para la ontología

3.4.2 Instancias

Para el prototipo de la bolsa de palabras, se decidió crear casos sobre el mismo para probar el sistema. Eso significa que se incluyeron algunas instancias en la ontología, por lo que la información en este momento está centralizada. Posteriormente la información será distribuida, tal que la ontología esté separada de la base de conocimiento. Los casos son necesarios para obtener algunos resultados de búsqueda con la bolsa de palabras. Hasta el momento, en la ontología se contemplan medicamentos para diversos tipos de enfermedades y síntomas.

3.4.3 Descripción de la ontología de clasificación de medicamentos

En la primera tarea se identificaron todos los términos, es decir los conceptos, relaciones, instancias y atributos, entre otros, que son importantes en el dominio, además, se enlistaron y se describieron en lenguaje natural.

3.4.3.1 Tarea 1: Glosario de términos

A continuación, se presentan las definiciones de los términos con su descripción, así como los conceptos ontológicos de domino de medicamentos y la industria farmacéutica, los cuales se observan en la Tabla 3.1.

Tabla 3. 1 Glosario de términos

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
FARMACÉUTICA	Farmacéutica	--	Concepto	Empresa que legalmente fabrica medicamentos para la venta en el mercado al por mayor o al por menor.
FÁRMACO	Fármaco, principio activo	--	Concepto	Conjunto de sustancias que producen un efecto farmacológico.
EFECTOS ADVERSOS	Reacción adversa a medicamentos	RAM	Concepto	Reacción provocada al consumir un medicamento.
ACNÉ	Acné	--	Concepto	Afección cutánea que se produce cuando los folículos pilosos se llenan de

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				grasa y células muertas de la piel.
ALOPECIA	Alopecia	--	Concepto	Caída del pelo en el cuero cabelludo o en cualquier otra parte del cuerpo.
ALTERACIONES CORNEALES	Alteraciones corneales, Queratitis	--	Concepto	Es una inflamación que afecta a la córnea, es decir la porción anterior y transparente del ojo.
ANAFILAXIA	Anafilaxia	--	Concepto	Reacción alérgica aguda y potencialmente mortal.
ANGIOEDEMA	Angioedema o edema de Quincke	--	Concepto	Reacción alérgica aguda y potencialmente mortal.
ARRITMIA CARDÍACA	Arritmia cardíaca, trastorno del ritmo cardíaco	--	Concepto	Latidos anormales del corazón, ya sea irregulares, demasiado rápidos o demasiado lentos.
BLOQUEO AURICULOVENTRICULAR	Bloqueo auriculoventricular, bloqueo AV	--	Concepto	Conjunto de trastornos del sistema de conducción que provocan que el estímulo eléctrico auricular sea conducido con retraso o no sea conducido a los ventrículos.
BRONCOESPASMO	Broncoespasmo	--	Concepto	Contracción anormal del músculo liso de los bronquios que provoca un estrechamiento u obstrucción aguda de las vías respiratorias.
CATARATAS	Cataratas	--	Concepto	Opacidad del cristalino de los ojos que normalmente es transparente.

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
CEFALEA	Cefalea, dolor de cabeza	--	Concepto	Dolor o incomodidad en la cabeza o la cara.
CISTITIS HEMORRÁGICA	Cistitis hemorrágica	--	Concepto	Es la aparición repentina de hematuria en combinación con dolor vesical y síntomas de irritación vesical. Se denomina hematuria a la presencia de sangre en la orina.
COLESTASIS	Colestasis, colestasia	--	Concepto	Afección en la que el flujo de bilis del hígado se hace más lento o se detiene por completo.
CONVULSIONES	Convulsiones	--	Concepto	Una contracción anormal e involuntaria de los músculos más comúnmente observada con ciertos trastornos convulsivos.
DELIRIOS	Delirios	--	Concepto	un estado de incapacidad para pensar o hablar claramente debido a fiebre o confusión mental.
DEPRESIÓN	Depresión	--	Concepto	Pérdida de interés en las actividades, lo que normalmente causa dificultades significativas en la vida cotidiana.
DERMATITIS LIQUENOIDE	Dermatitis liquenoide	--	Concepto	Alergia que se forma en la piel y causa una erupción púrpura. Es causada por el daño a la epidermis, lo que resulta en la inflamación y los problemas con la forma en la que capa

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				externa de la piel interactúa con la siguiente capa interna de tejido, llamada dermis.
DIARREA	Diarrea	--	Concepto	Enfermedad caracterizada por deposiciones flojas y acuosas o una necesidad frecuente de evacuar.
DISFUNCIÓN RENAL	Disfunción renal, insuficiencia renal	--	Concepto	Afección que provoca que los riñones pierdan la capacidad de eliminar los desechos y equilibrar los fluidos.
DISFUNCIÓN SEXUAL	Disfunción sexual	--	Concepto	Dificultad durante cualquier etapa del acto sexual (que incluye deseo, excitación, orgasmo y resolución) que evita al individuo o pareja el disfrute de la actividad sexual.
ECZEMA	eczema, eccema, dermatitis eccematosa	--	Concepto	Conjunto de afecciones dermatológicas, caracterizadas por presentar lesiones inflamatorias.
EDEMA PULMONAR	Edema pulmonar	--	Concepto	Es una afección que se produce cuando los pulmones se llenan de líquido. Cuando sucede esto, se ve afectada la capacidad del organismo de recibir la cantidad suficiente de oxígeno.
ENFERMEDAD DEL SUERO	Enfermedad del suero	--	Concepto	Conjunto de síntomas sistémicos

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				reversibles provocado por la formación de inmunocomplejos formados a partir de antígenos extraños presentes en la circulación.
EPIDERMOLISIS	Epidermolisis	--	Concepto	Enfermedad que hace que la piel se vuelva frágil formándose ampollas dolorosas.
ERITEMA NODOSO	Eritema nodoso	--	Concepto	Lesión de la piel que consiste en la presencia de nódulos dolorosos que presentan signos inflamatorios (enrojecimiento y aumento de temperatura) y se localizan predominantemente a la zona pretibial (delante de la tibia), en las extremidades inferiores, aunque es posible que se encuentre en otras localizaciones.
ESTREÑIMIENTO	Estreñimiento	--	Concepto	Alteración del intestino que hace difícil la expulsión de las heces.
FIEBRE	Fiebre	--	Concepto	Incremento de la temperatura corporal
FOTODERMATITIS	Fotodermatitis	--	Concepto	Inflamación de la piel en la que la acción de la luz es el mecanismo desencadenante.
GALACTORREA	Galactorrea	--	Concepto	Es una afección que se produce cuando uno o ambos senos producen leche (o

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				una secreción lechosa).
GINECOMASTIA	Ginecomastia	--	Concepto	Inflamación del tejido mamario masculino provocada por un desequilibrio hormonal.
GLAUCOMA	Glaucoma	--	Concepto	Conjunto de afecciones oculares que posiblemente provocan la ceguera.
HEPATITIS	Hepatitis	--	Concepto	Inflamación del hígado.
HIPERTENSIÓN	Hipertensión	--	Concepto	Afección en la que la presión de la sangre hacia las paredes de la arteria es demasiado alta.
HIPERTENSIÓN INTRACRANEAL	Hipertensión intracraneal	--	Concepto	Incremento en la presión hidrostática del interior de la cavidad craneal, en particular en el líquido cefalorraquídeo, debido a la suma de presiones que ejercen los elementos intracraneales.
HIPOTENSIÓN	Hipotensión	--		Presión excesivamente baja de la sangre sobre la pared de las arterias.
LITIASIS RENAL	Litiasis renal, Urolitiasis, nefrolitiasis	--	Concepto	Pequeños depósitos endurecidos que se forman en los riñones y suelen doler al ser expulsados.
LUPUS ERITEMATOSO	Lupus eritematoso, lupus eritematoso sistémico	LES	Concepto	Enfermedad autoinmune inflamatoria crónica del tejido conjuntivo (conectivopatía) poco habitual que afecta a todo el

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				organismo y con múltiples manifestaciones clínicas desde la piel, las articulaciones, los riñones, el corazón, el cerebro o hasta alteraciones psicológicas (depresión, ansiedad).
MALABSORCIÓN	Malabsorción, síndrome de malabsorción	--	Concepto	Afección que impide la absorción de nutrientes a través del intestino delgado.
MANÍAS	Manías	--	Concepto	Es un trastorno mental consistente en una elevación anómala del estado anímico. Forma parte de los trastornos del ánimo, constituyendo una de las fases del llamado trastorno bipolar.
MIOCARDIOPATÍA	Miocardiopatía	--	Concepto	Enfermedad hereditaria o adquirida del miocardio.
NECROSIS TUBULAR	Necrosis tubular, necrosis tubular aguda	NTA	Concepto	La necrosis tubular aguda es la lesión renal causada por el daño a las células de los túbulos renales (células renales que reabsorben los líquidos y minerales de la orina a medida que va generándose esta).
NEFRITIS INTERSTICIAL	Nefritis intersticial, nefritis tubulointersticia	--	Concepto	Es un trastorno renal en el cual los espacios entre los túbulos renales

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
	l o nefritis intersticial aguda			resultan hinchados (inflamados). Es posible que la inflamación afecte el funcionamiento de los riñones, incluyendo su capacidad para filtrar los desechos.
NEURITIS ÓPTICA	Neuritis óptica	--	Concepto	Inflamación del nervio óptico.
NEUROPATÍA PERIFÉRICA	Neuropatía periférica	--	Concepto	Debilidad, entumecimiento y dolor, generalmente en las manos y los pies, ocasionado por un daño neurológico.
NÁUSEAS	Náuseas	--	Concepto	Sensación de malestar en el estómago que se experimenta cuando se tienen ganas de vomitar y que suele culminar en vómitos.
PANCREATITIS	Pancreatitis	--	Concepto	Inflamación del órgano ubicado detrás de la parte inferior del estómago (páncreas).
PERICARDITIS	Pericarditis	--	Concepto	Inflamación y también irritación de la delgada membrana similar a un saco que rodea al corazón (pericardio).
PÚRPURA	Púrpura, síndrome purpúrico	--	Concepto	Enfermedad que hace a los humanos sensibles a la luz y conduce a la acumulación anormal de porfirinas
RETINOPATÍAS	Retinopatías	--	Concepto	Daño en los vasos sanguíneos del tejido ubicado en la parte

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				posterior del ojo (retina).
SORDERA	Sordera	--	Concepto	Es la dificultad o la imposibilidad de usar el sentido del oído debido a una pérdida de la capacidad auditiva parcial (hipoacusia) o total (cofosis), y unilateral o bilateral.
SÍNDROME DE STEVENS-JOHNSON	Síndrome de Stevens-Johnson	SSJ	Concepto	Es una enfermedad rara y grave de la piel y las membranas mucosas. SSJ le hará perder hasta el 10% de la capa externa de la piel.
SÍNDROME NEFRÓTICO	Síndrome nefrótico	SN	Concepto	Enfermedad del riñón que ocasiona que el cuerpo elimine demasiadas proteínas en la orina.
TOS	Tos	--	Concepto	Reacción natural que permite al organismo proteger los pulmones y las vías respiratorias contra partículas extrañas y contra la acumulación de flemas y moco, evitando así que éstas se obstruyan.
TRASTORNO DEL SUEÑO	Trastorno del sueño	--	Concepto	Cambios en los patrones o hábitos del sueño que afectan a la salud.
TROMBOEMBOLISMO	Tromboembolismo	--	Concepto	Es una enfermedad que incluye trombosis venosa profunda (TVP) y embolia pulmonar (EP)
ULCERA PÉPTICA	Ulcera péptica	--	Concepto	Llaga que aparece en el recubrimiento del esófago, el estómago

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				o el intestino delgado.
URTICARIA	Urticaria	--	Concepto	Sarpullido en la piel ocasionado por una reacción a la comida, los medicamentos y otros agentes irritantes.
VÓMITO	Vómito	--	Concepto	Expulsión violenta por la boca de lo que está contenido en el estómago.
MEDICAMENTO	Medicamento	--	Concepto	Sustancia con potencial para prevenir o curar una enfermedad
ANALGÉSICOS	Analgésicos	--	Concepto	Medicamento diseñado específicamente para aliviar el dolor.
ANALGÉSICOS NO OPIOIDES	Analgésicos no opioides	--	Concepto	Los analgésicos no opioides son principalmente empleados para tratar el dolor de moderado a intenso.
ANALGÉSICOS OPIOIDES	Analgésicos opioides	--	Concepto	Los analgésicos opioides son un grupo de fármacos que poseen gran actividad analgésica, mediada por la activación de receptores específicos en el sistema nervioso central y periférico, que son derivados naturales del opio, o bien derivados semisintéticos o análogos sintéticos con muchas características en común.
ANTIMIGRAÑOSOS	Antimigrañosos	--	Concepto	Medicamentos que alivian el dolor de

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				cabeza o disminuyen la frecuencia de los episodios de migraña.
ANTIPIRÉTICOS	Antipiréticos	--	Concepto	Medicamentos que tratan la fiebre de una forma sintomática, sin actuar sobre su causa.
ANESTÉSICOS	Anestésicos	--	Concepto	Produce la pérdida temporal del conocimiento o de la sensibilidad de una parte del cuerpo.
ANTIAGREGANTES	Antiagregantes	--	Concepto	Los antiagregantes plaquetarios impiden la formación de coágulos sanguíneos en las arterias. Los coágulos en las arterias del corazón y el cerebro son causales de un infarto de miocardio o un ictus.
ANTIANÉMICOS	Antianémicos	--	Concepto	Fármacos que favorecen la eritropóyesis incrementando la formación de eritrocitos y/o hemoglobina.
ANTIASMÁTICOS	Antiasmáticos	--		Calma o atenúa los síntomas del asma.
ANTIBIÓTICOS	Antibióticos	--	Concepto	Son fármacos usados en el tratamiento de infecciones por bacterias.
ANTICOAGULANTES	Anticoagulantes	--	Concepto	Sirve para inhibir o retardar la coagulación de la sangre.
ANTIDEPRESIVOS	Antidepresivos	--	Concepto	Medicamentos utilizados para el tratamiento del trastorno depresivo

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				mayor y otras afecciones.
ANTIDIABÉTICOS	Antidiabéticos	--	Concepto	Previene o combate la diabetes.
ANTIDIABÉTICOS ORALES SULFONILUREAS	Antidiabéticos orales sulfonilureas, sulfonilureas	--	Concepto	Las sulfonilureas son medicamentos pertenecientes a la clase de antidiabéticos orales indicados en el tratamiento de la diabetes mellitus tipo 2. Actúan aumentando la liberación de insulina de las células beta del páncreas.
INSULINAS HUMANAS	Insulinas humanas	--	Concepto	Está indicada en el tratamiento de la diabetes mellitus, cuando se precise tratamiento con insulina. También está indicada para el tratamiento del coma hiperglucémico y de la cetoacidosis diabética, así como para obtener la estabilización pre, intra y postoperatoria de pacientes con diabetes mellitus.
INSULINA INTERMEDIA	Insulina intermedia	--	Concepto	Se absorbe más lentamente, y dura más. Se usa para controlar el azúcar en sangre durante la noche, mientras se está en ayunas y entre comidas.
INSULINA LENTA	Insulina lenta, insulinas basales	--	Concepto	Las insulinas basales son aquellas que se inyectan una o dos veces al día.

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
INSULINA ULTRARRÁPIDA	Insulina ultrarrápida	--	Concepto	Es una insulina transparente. Es recomendable administrarla durante o poco después de las comidas.
ANTIDIARREICOS	Antidiarreicos	--	Concepto	Medicamentos que sirven para combatir la diarrea.
ANTIEMÉTICOS	Antieméticos	--	Concepto	Impide o evita el vómito.
ANTIEPILÉPTICOS	Antiepilépticos	--	Concepto	Substancia destinada a combatir, prevenir o interrumpir las convulsiones o los ataques epilépticos.
ANTIESPASMÓDICOS	Antiespasmódicos	--	Concepto	Se utiliza para aliviar o calmar los espasmos o convulsiones.
ANTIFLATULENTOS	Antiflatulentos	--	Concepto	Se utilizan para el alivio o la prevención del exceso de gases intestinales (es decir, las flatulencias).
ANTIGOTOSOS	Antigotosos	--	Concepto	Son medicamentos que controlan el ácido úrico para prevenir o tratar los tofos (masas nodulares de cristales de ácido úrico).
ANTIHEMORROIDALES	Antihemorroidales	--	Concepto	Ayudan a aliviar y a su vez controlar las molestias producidas por las hemorroides.
ANTIHIPERAMONÉMICOS	Antihiperamonémicos	--	Concepto	Reducen el pH en el colon y en consecuencia reducen la absorción de amoníaco y aumentan la eliminación de NH4+ en pacientes hiperamonémicos.

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
ANTIHIPERTENSIVOS	Antihipertensivos	--	Concepto	Los antihipertensivos son un grupo de medicamentos utilizados para normalizar una tensión arterial anormalmente alta.
ANTIHISTAMÍNICOS	Antihistamínicos	--	Concepto	Alivia o previene los síntomas de alergia.
ANTIINFLAMATORIOS	Antiiflamatorios	--	Concepto	Reducen o combaten los síntomas y los signos de la inflamación.
ANTIINFLAMATORIOS NO ESTEROIDES	Antiinflamatorios no esteroides	AINE	Concepto	Los fármacos antiinflamatorios no esteroides o AINE son medicamentos que se usan para tratar tanto el dolor como la inflamación.
ANTIMICÓTICOS	Antimicóticos	--	Concepto	Los medicamentos antimicóticos se usan para tratar infecciones por hongos como la candidiasis, la tiña o el pie de atleta.
ANTIPARASITARIOS	Antiparasitarios	--	Concepto	Utilizado o destinado a matar, repeler o eliminar parásitos.
ANTIPARKINSONIANOS	Antiparkinsonianos	--	Concepto	Medicamentos útiles para el tratamiento de la enfermedad de Parkinson.
ANTIPSORIÁSICOS	Antipsoriásicos	--	Concepto	Producto que alivia los síntomas de la psoriasis.
ANTISÉPTICO	Antiséptico	--	Concepto	Son sustancias antimicrobianas que se aplican a un tejido vivo o sobre la piel para reducir la posibilidad de infección, sepsis o putrefacción.

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
ANTITUBERCULOSOS	Antituberculosos	--	Concepto	Medicamentos contra la tuberculosis.
ANTITUSÍGENOS	Antitusígenos, antitusivo	--	Concepto	Empleado para tratar la tos seca irritativa, no productiva. Los antitusígenos son compuestos que actúan sobre el sistema nervioso central o periférico para suprimir el reflejo de la tos.
ANTIULCEROSOS	Antiulcerosos	--	Concepto	Medicamento que cura o facilita la cicatrización de una úlcera o lesión abierta de la piel o membrana mucosa con forma de cráter y escasa o nula tendencia a la cicatrización espontánea.
ANTIVARICOSOS	Antivaricosos	--	Concepto	Ayuda a controlar los síntomas producidos por las várices (dilatación permanente de las venas) que comúnmente se presentan en las piernas, entre ellos, dolor, pesadez y calambres.
ANTIVERTIGINOSOS	Antivertiginosos	--	Concepto	Medicamento para prevenir o aliviar los síntomas de vértigo (sensación de movimiento o giro del entorno o de uno mismo), como mareo, náuseas o pérdida del equilibrio.
ANTIVIRALES	Antivirales	--	Concepto	Medicamento usado para el tratamiento

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				de infecciones producidas por virus. Tal como los antibióticos (específicos para bacteria), existen antivirales específicos para distintos tipos de virus.
ANTIÁCIDOS	Antiácidos	--	Concepto	Es una sustancia, generalmente una base, que actúa a favor de la acidez estomacal.
ANTÍDOTOS	Antídotos	--	Concepto	Sustancia química cuya función es contrarrestar los efectos de un veneno, toxina o químico.
CICATRIZANTE	Cicatrizante	--	Concepto	Favorece y acelera la cicatrización.
CORTICOIDES	Corticoides	--	Concepto	Son un tipo de las hormonas denominadas esteroides.
DESCONGESTIONANTES	Descongestionantes	--	Concepto	Evita la congestión nasal. Generalmente, trabajan reduciendo la hinchazón de las membranas mucosas en los pasos nasales.
DIURÉTICOS	Diuréticos	--	Concepto	Facilitan o aumentan la eliminación de orina.
EMOLIENTES	Emolientes	--	Concepto	Es un medicamento de uso externo que tiene la propiedad de ablandar una parte inflamada.
ENZIMAS DIGESTIVOS	Enzimas digestivos	--	Concepto	son medicamentos que se usan para aliviar la indigestión, la reducción de la

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
				inflamación, entre otros.
ESCLEROSANTES	Esclerosantes	--	Concepto	Se utilizan para el tratamiento de las malformaciones vasculares.
FIBRINOLÍTICOS	Fibrinolíticos	--	Concepto	Producen la destrucción de algunos factores de la coagulación de la sangre de una forma rápida, de tal manera que los coágulos recién formados se suelen deshacer, y el paso de la sangre se restablece.
HEMOSTÁTICOS	Hemostáticos	--	Concepto	Son medicamentos destinados a controlar y detener las hemorragias, tanto internas como externas.
HIPOLIPEMIANTES	Hipolipemiantes	--	Concepto	Disminuye los niveles de lípidos en sangre.
HORMONAS SEXUALES	Hormonas sexuales	--	Concepto	Son las encargadas de regular el ciclo reproductivo en mujeres y hombres.
INMUNOMODULADORES	Inmunomoduladores	--	Concepto	Sustancia que estimula o deprime el sistema inmunitario, y ayuda al cuerpo a combatir el cáncer, las infecciones u otras enfermedades.
INMUNOSUPRESORES	Inmunosupresores	--	Concepto	Son medicamentos que reducen o bloquean las reacciones inmunitarias del organismo.
LAXANTES	Laxantes	--	Concepto	Facilita la expulsión de los excrementos del intestino.

NOMBRE	SINÓNIMO	ACRÓNIMO	TIPO	DESCRIPCIÓN
MIORRELAJANTES	Miorrelajantes	--	Concepto	Efectivos para el alivio sintomático a corto plazo en los pacientes con dolor lumbar agudo y crónico.
MUCOLÍTICOS	Mucolíticos	--	Concepto	Sirven como auxiliares para el alivio de los síntomas de las infecciones respiratorias.
OFTALMOLÓGICOS	Oftalmológicos	--	Concepto	Se utilizan para tratar la patología de los ojos. Algunos son de uso exclusivamente oftálmico y otros son de uso general o preparaciones especiales de uso general específicas para los ojos.
PSICOLÉPTICOS	Psicolépticos	--	Concepto	Son sustancias cuyos efectos psicotrópicos aumentan la actividad del sistema nervioso. Estos estimulantes psíquicos excitan el humor.
FABRICAMEDICAMENTO	fabricaMedicamento	--	Relación	Asociación entre Farmacéutica y Medicamento
ESFABRICADOPOR	esFabricadoPor	--	Relación	Asociación entre Medicamento y Farmacéutica .
CAUSAEFECTOSADVERSOS	causaEfectosAdversos	--	Relación	Asociación entre Medicamento y Efectos adversos
ESCAUSADOPOR	esCausadoPor	--	Relación	Asociación entre Efectos adversos y Medicamento
CONTIENEFARMACO	contieneFarmaco	--	Relación	Asociación entre Medicamento y Farmaco
ESCONTENIDOEN	esContenidoEn	--	Relación	Asociación entre Farmaco y Medicamento

3.4.3.2 Tarea 2: Taxonomía de conceptos

La taxonomía de conceptos trata los principios y métodos para la clasificación y organización jerarquizada y sistemática para asignar nombres a grupos de elementos y a los elementos propios de la taxonomía de medicamentos y fármacos de la bolsa de palabras. En este sentido, se estableció una jerarquía entre los conceptos del dominio basándose en el glosario de términos, que se muestra en la Figura 3.21.

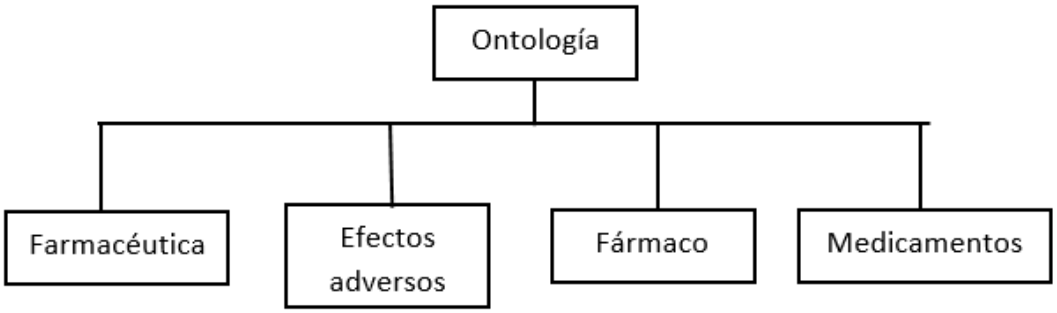

Figura 3. 21 Taxonomía general (parte 1)

En la Figura 3.22 se muestra la taxonomía general dividida en cuatro términos como se mostró anteriormente, donde cada término tiene asignados diversos términos correspondientes al glosario.

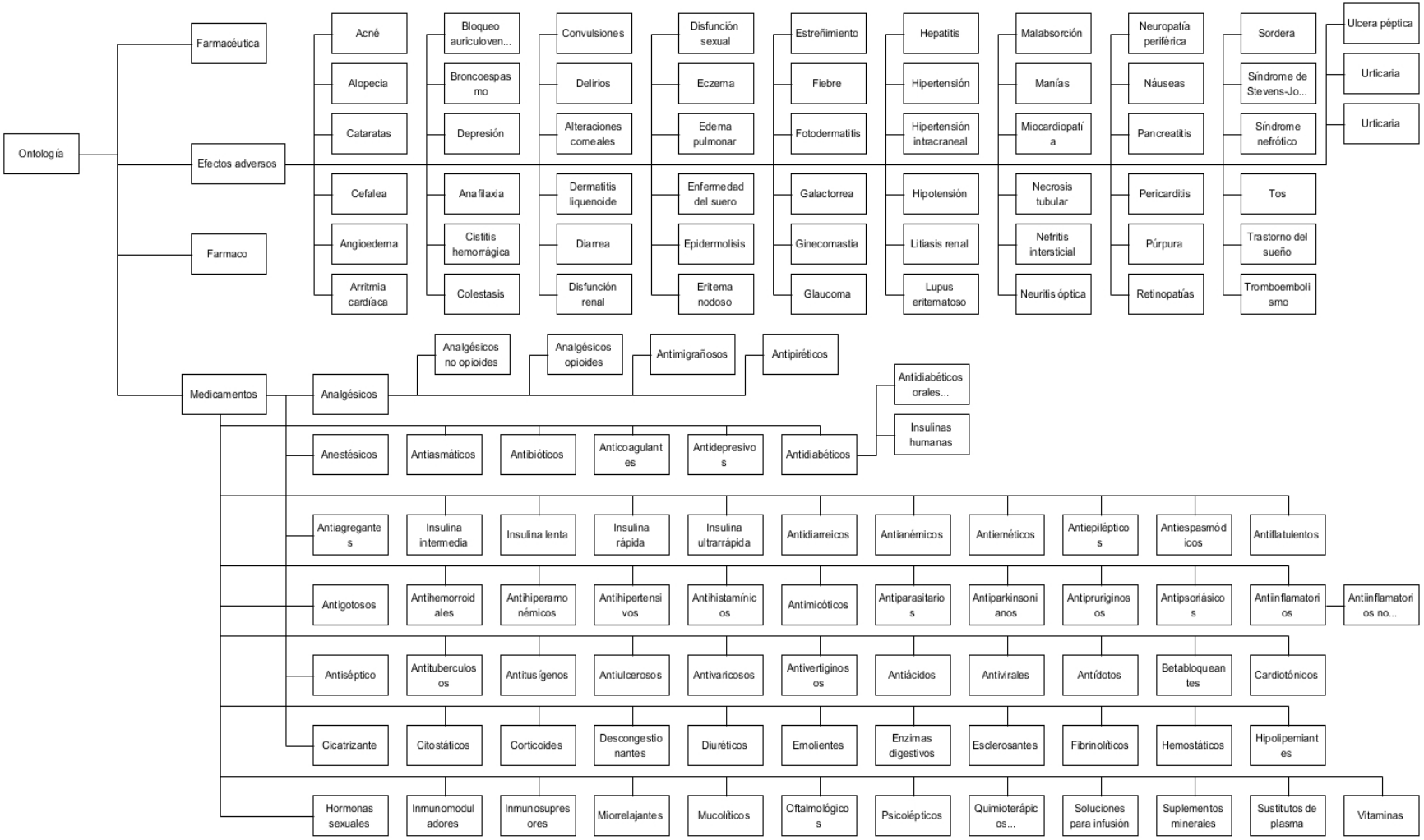

Figura 3. 22 Taxonomía general (parte 2)

3.4.3.3 Tarea 3: Diagrama de relaciones binarias

En esta tarea se construyó un diagrama que representa las relaciones directas e indirectas existentes entre los conceptos como se muestra en la Figura 3.23 y en la Figura 3.24.

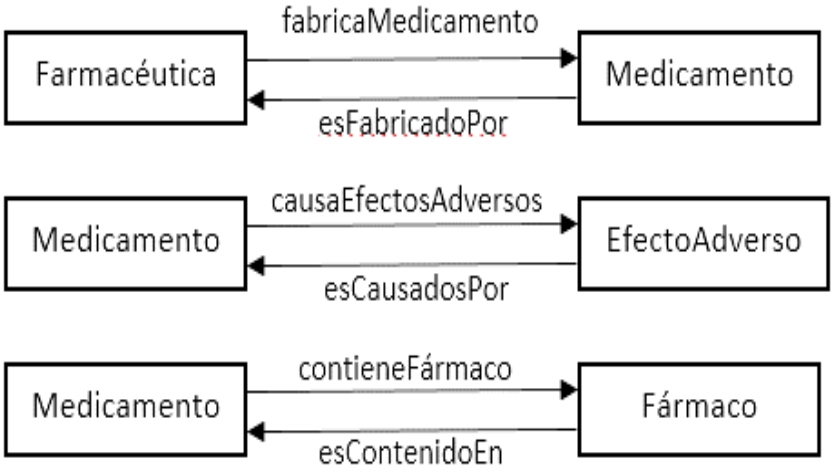

Figura 3. 23 Mapa conceptual de las relaciones (parte 1)

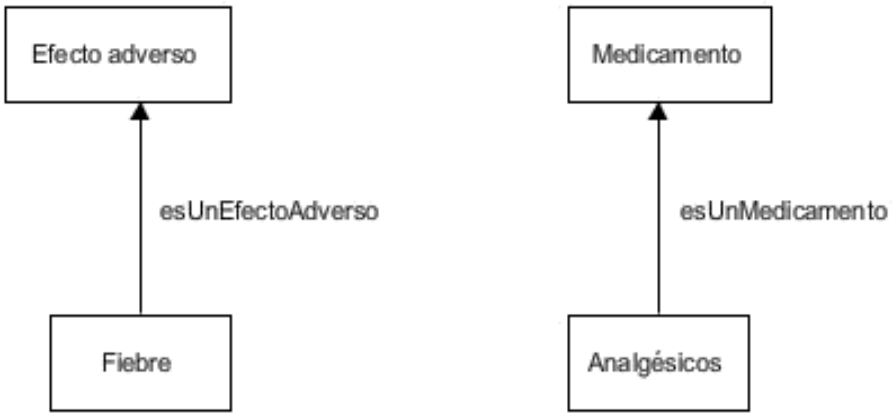

Figura 3. 24 Mapa conceptual de las relaciones (parte 2)

A continuación, se representan de una en la Tabla 3.2.

Tabla 3. 2 Relaciones entre conceptos

NOMBRE	CONCEPTOA	CONCEPTOB	TIPO
FABRICAMEDICAMENTO	Farmacéutica	Medicamento	No simétrica
ESFABRICADOPOR	Medicamento	Farmacéutica	No simétrica
CAUSAEFECTOSADVERSOS	Medicamento	Efectos adversos	No simétrica
ESCAUSADOPOR	Efectos adversos	Medicamento	No simétrica
CONTIENEFARMACO	Medicamento	Fármaco	No simétrica
ESCONTENIDOEN	Fármaco	Medicamento	No simétrica

3.4.3.4 Tarea 4: Diccionario de conceptos de mayor relevancia

Se identificaron los conceptos más relevantes del dominio dando como resultado una lista de conceptos y las relaciones existentes.

Tabla 3. 3 Conceptos de mayor relevancia

CONCEPTO	RELACIONES
FARMACÉUTICA	fabricaMedicamento → Medicamento
FÁRMACO	esContenidoEn → Medicamento
EFECTOS ADVERSOS	esCausadoPor → Medicamento
MEDICAMENTO	esFabricadoPor → Farmacéutica causaEfectosAdversos → Efectos adversos contieneFarmaco → Fármaco

La Figura 3.25 muestra los conceptos de mayor relevancia de la ontología de medicamentos para la bolsa de palabras diseñada con la herramienta Protégé en su primer nivel, es decir, con sus sub-conceptos contraídos.

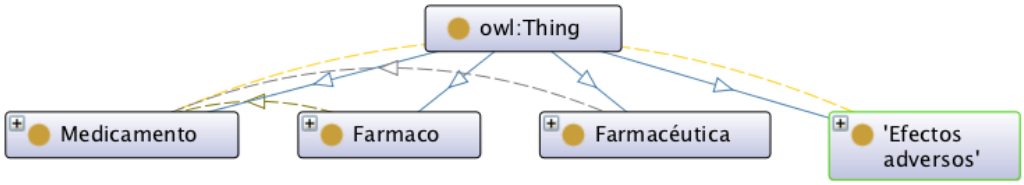

Figura 3. 25 Conceptos de mayor relevancia

3.4.3.5 Tarea 5: Descripción de relaciones binarias como dominio

Se identificaron los detalles de las relaciones como dominio, rango y cardinalidad dando como resultado una lista de relaciones, concepto origen, concepto destino y cardinalidad, visibles en la Tabla 3.4.

Tabla 3. 4 Identificación de relaciones como dominio

RELACIÓN	DOMINIO	RANGO	CARDINALIDAD
FABRICAMEDICAMENTO	Farmacéutica	Medicamento	1:*…1:*
ESFABRICADOPOR	Medicamento	Farmacéutica	1:*…1:*
CAUSAEFECTOSADVERSOS	Medicamento	Efectos adversos	1:*…1:*
ESCAUSADOPOR	Efectos adversos	Medicamento	1:*…1:*
CONTIENEFARMACO	Medicamento	Fármaco	1:*…1:*
ESCONTENIDOEN	Fármaco	Medicamento	1:*…1:*

3.4.3.6 Tarea 6: Identificación de atributos de las instancias

Se identificaron los atributos que cualquier instancia de un concepto tiene, dando como resultado una lista de atributos, concepto al que pertenecen, tipo de valor, rango de valores y cardinalidad, como se muestra en la Tabla 3.5.

Tabla 3. 5 Identificación de atributos e las instancias

ATRIBUTO	CONCEPTO	TIPO DE VALOR	RANGO DE VALORES	CARDINALIDAD
NOMBRE	Medicamento	Cadena	Mínimo 2 caracteres	1:1
VÍA DE ADMINISTRACIÓN	Medicamento	Cadena	Mínimo 4 letras	1:1
PRESENTACIÓN	Medicamento	Cadena	Mínimo 6 letras	1:1

ATRIBUTO	CONCEPTO	TIPO DE VALOR	RANGO DE VALORES	CARDINALIDAD
DOSIS	Medicamento	Cadena	Mínimo 20 caracteres	1:1
INDICACIONES	Medicamento	Cadena	Mínimo 20 caracteres	1:1
CONTRAINDICACIONES	Medicamento	Cadena	Mínimo 20 caracteres	1:1
PRECAUCIONES	Medicamento	Cadena	Mínimo 20 caracteres	1:1
NOMBRE FARMACÉUTICA	Farmacéutica	Cadena	Mínimo 3 caracteres	1:1
DESCRIPCIÓN FARMACÉUTICA	Farmacéutica	Cadena	Mínimo 8 caracteres	1:1
DIRECCIÓN FARMACÉUTICA	Farmacéutica	Cadena	Mínimo 20 caracteres	1:1
TELÉFONO FARMACÉUTICA	Farmacéutica	Numérico	Mínimo 7 digitos	1:1
NOMBRE FÁRMACO	Fármaco	Cadena	Mínimo 3 caracteres	1:1
DESCRIPCIÓN FÁRMACO	Fármaco	Cadena	Mínimo 10 caracteres	1:1
FÓRMULA QUÍMICA	Fármaco	Cadena	Mínimo 6 caracteres	1:1

3.4.3.7 Tarea 7: Descripción de atributos de clase

En esta tarea se especifican los atributos pertenecientes a cada clase, el valor y la cardinalidad, sin embargo, esta tarea no se aplica para la ontología de dominio médico que se emplean en la presente tesis.

3.4.3.8 Tarea 8: Descripción de constantes

El objetivo de esta tarea es ofrecer una descripción de las constantes identificadas para la ontología mediante la especificación de su nombre, valor, unidad de medida (en el caso de las constantes numéricas). Esta tarea no aplica para la ontología de dominio médico que se emplea.

3.4.3.9 Tarea 9: Descripción de axiomas formales

En esta tarea se identificaron los axiomas que representan restricciones del dominio y se determinan sus características, como se muestra en la Tabla 3.6.

Tabla 3. 6 Identificación de axiomas y restricciones

NOMBRE	DESCRIPCIÓN	EXPRESIÓN	CONCEPTOS	RELACIO-NES	VARIA-BLES
ACNÉ ES UN EFECTO ADVERSO	Acné es un Efecto adverso	SubClassOf(a:Acné a:Efecto adverso)	Acné, Efecto adverso	SubClassOf	a
TRASTORNO DEL SUEÑO ES UN EFECTO ADVERSO	Cada Trastorno del sueño es un Efecto adverso	SubClassOf(a:Trastorno del sueño a:Efecto adverso)	Trastorno del sueño, Efecto adverso	SubClassOf	a
UN ANALGÉSICO ES UN MEDICAMENTO	Cada Analgésico es un Medicamento	SubClassOf(a:Analgesico a:Medicamento)	Analgesico, Medicamento	SubClassOf	a
UN ANTIMIGRAÑOSO ES UN ANALGÉSICO	Cada Antimigrañoso es un Analgésico	SubClassOf(a:Antimigrañoso a:Analgésico)	Antimigrañoso, Analgésico	SubClassOf	a
UN ANTIBIÓTICO ES UN MEDICAMENTO	Cada Antibiótico es un Medicamento	SubClassOf(a:Antibiótico a:Medicamento)	Antibiótico, Medicamento	SubClassOf	a
UN ANTIINFLAMATORIO ES UN MEDICAMENTO	Cada Antiinflamatorio es un Medicamento	SubClassOf(a:Antiinflamatorio a:Medicamento)	Antiinflamatorio, Medicamento	SubClassOf	a
UN ANTIINFLAMATORIO NO ESTEROIDE ES UN ANTIINFLAMATORIO	Cada Antiinflamatorio no esteroide es un Antiinflamatorio	SubClassOf(a:Antiinflamatorio no esteroide a:Antiinflamatorio)	Antiinflamatorio no esteroide, Antiinflamatorio	SubClassOf	a
INCOMPATIBILIDAD FARMACÉUTICA Y FÁRMACO	Nada puede ser una Farmacéutica y un Fármaco	DisjointClasses(a:Farmacéutica a:Farmaco)	Farmacéutica, Farmaco	DisjointClasses	a
INCOMPATIBILIDAD EFECTO ADVERSO Y MEDICAMENTO	Nada puede ser un Efecto adverso y un Medicamento	DisjointClasses(a:Efecto adverso a:Medicamento)	Efecto adverso, Medicamento	DisjointClasses	a
INCOMPATIBILIDAD FARMACÉUTICA	Nada puede ser una Farmacéutica y	DisjointClasses(a:Farmacéutica a:Medicamento)	Farmacéutica, Medicamento	DisjointClasses	a

NOMBRE	DESCRIPCIÓN	EXPRESIÓN	CONCEPTOS	RELACIO-NES	VARIA-BLES
Y MEDICAMENTO	un Medicamento				
INCOMPATIBILIDAD EFECTO ADVERSO Y FÁRMACO	Nada puede ser un Efecto adverso y un Fármaco	DisjointClasses(a:Efecto adverso a:Farmaco)	Efecto adverso, Farmaco	DisjointClasses	a

3.4.3.10 Tarea 10: Descripción de reglas formales

Se identificaron y describieron las reglas del dominio dando como resultado una lista de reglas y sus características.

Tabla 3. 7 Identificación de reglas de dominio

NOMBRE	DESCRIPCIÓN	EXPRESIÓN	CONCEPTOS/RELACIONES
ENCONTRAR INVERSA DE CAUSAEFECTOSADVERSOS	Si un efecto adverso es ocasionado por un medicamento entonces el medicamento causaEfectosAdversos Efectos adversos.	Si medicamento(?x) AND efectos adveros(?y) AND causaEfectosAdversos (?x, ?y) Entonces esCausadoPor (?y, ?x)	Medicamento, efectosAdversos, causaEfectosAdversos, esCausadoPor
ENCONTRAR INVERSA DE FABRICAMEDICAMENTO	Si un medicamento es fabricado por una farmacéutica, entonces la farmacéutica fabricaMedicamento medicamento.	Si farmacéutica(?x) AND medicamento(?y) AND fabricaMedicamento (?x, ?y) Entonces esFabricadoPor (?y, ?x)	Farmacéutica, mediamento, fabricaMedicamento, esFabricadoPor

3.4.3.11 Tarea 11: Descripción de las instancias del dominio

Se describieron algunos ejemplos de instancias del dominio. Cada instancia cuenta con nombre, nombre del concepto al que pertenece y los valores de sus atributos de instancia, cuando estos se conocían.

Tabla 3. 8 Instancias del dominio

NOMBRE	CONCEPTO	VALORES DE ATRIBUTOS
ASPIRINA 500 MG	Medicamento	Nombre: Aspirina 500 mg, vía de administración: oral, presentación: sobres, dosis: disuelva el contenido del sobre en medio vaso de agua. Tome 1 sobre en el desayuno, en la comida y en la cena, indicaciones: medicamento que disminuye el dolor (analgésico), la inflamación (antiinflamatorio) y la fiebre (antipirético). También se usa para prevenir el infarto porque evita la formación de obstrucciones (coágulos) de sangre en las arterias, contraindicaciones: no tome este medicamento si es alérgico a ácido acetilsalicílico (aspirina), precauciones: No se recomienda el uso en influenza ni en varicela, ya que su uso se ha asociado al síndrome de Reye en niños menores de 14 años.
ÁCIDO ACETILSALICÍLICO	Fármaco	Nombre fármaco: Ácido acetilsalicílico, descripción fármaco: es un fármaco de la familia de los salicilatos. Se utiliza como medicamento para tratar el dolor (analgésico), la fiebre (antipirético) y la inflamación (antiinflamatorio), debido a su efecto inhibitorio, no selectivo, de la ciclooxigenasa, fórmula química: C9H8O4.

Capítulo 4. Resultados

Para mostrar el funcionamiento del enfoque planteado en esta investigación, se desarrolló el sistema que lleva por nombre "SentiScrap" bajo la arquitectura propuesta en capítulos previos. El sistema se encuentra bajo la modalidad de aplicación Web.

SentiScrap proporciona al usuario la capacidad de analizar foros que se encuentran en la Web y video blogs sobre medicamentos para enfermedades crónico-degenerativas, en concreto para las enfermedades: Diabetes, Hipertensión y Hepatitis. El sistema es capaz de realizar la extracción necesaria a dichas fuentes para obtener las opiniones de pacientes y especialistas de salud, lo que da como resultado un conjunto de información muy valiosa a analizar mediante el reconocimiento del dominio y la detección de polaridad.

La información obtenida se muestra en la aplicación Web, donde se muestra la información resultante de las tres enfermedades a las cuales el análisis tiene alcance.

4.1 Caso de estudio

4.1.1 Análisis de polaridad en comentarios publicados en blogs, foros y video blogs en el idioma español de medicamentos para enfermedades crónico-degenerativas con validación de un experto en la salud.

La comunidad médica ha detectado una falta de registros de pacientes que describan los síntomas de los Efectos Adversos a los Medicamentos (ADE's), incluso con aquellos no identificados previamente por las compañías farmacéuticas. Este hecho representa una desinformación por parte de los especialistas de la salud, así como de los propios pacientes, por no informar al médico en las consultas sobre los síntomas añadidos a su condición.

Las enfermedades crónico-degenerativas se destacan por posicionarse entre las 10 principales causas de muerte en México, lo que representa una gran cantidad de población que acude con su médico a consultas y chequeos rutinarios. Sin embargo, ante esa exigencia, los especialistas de la salud cuentan con poco tiempo para revisar a los pacientes en las consultas, en consecuencia, la exploración física resulta ser poco empleada por requerir más tiempo. Ante este hecho, se omiten nuevos síntomas independientes al padecimiento inicial, que

probablemente son causados por los fármacos o algún otro tipo de enfermedad independiente, que termina por ser mal tratada ante la falta de detección.

Para este caso de estudio, se supone que un especialista de la salud necesita conocer las opiniones de los pacientes que están bajo un tratamiento para tratar la Hipertensión:

- ¿Cómo podrá identificar el especialista los medicamentos de los que hablan los pacientes para el tratamiento de la Hipertensión?
- ¿Cómo conocerá el especialista los comentarios de los pacientes que publican en foros y video blogs en el idioma español?
- ¿De qué forma el especialista conocerá el impacto positivo y negativo en los comentarios?
- ¿Cómo podría el especialista validar bajo su experiencia médica los comentarios que comparten los pacientes?
- ¿De qué manera impactará al especialista conocer los comentarios realizados por pacientes?
- ¿Cómo contribuirá el especialista de la salud a incrementar el conocimiento de SentiScrap?

El especialista de la salud tendrá acceso a la información que provee la aplicación Web, sin tener que hacer un registro previo de inicio de sesión, ya que la información contenida en SentiScrap tiene como objetivo estar al alcance de todos los especialistas de la salud.

El sistema provee un menú con opciones de tres tipos de enfermedades sobre las que revisa la información analizada, además una opción para añadir fuentes y una opción para revisar gráficas relacionadas con el análisis de polaridad, en Figura 4.1se muestra el contenido de la pantalla principal.

Figura 4. 1 Inicio de SentiScrap

Suponiendo que un especialista de la salud quiere conocer aquellos medicamentos que se prescriben para el tratamiento de la hipertensión, en cuyo caso el especialista selecciona la enfermedad "Hipertensión" que se encuentra en el menú, en consecuencia, el sistema genera una consulta al repositorio para mostrar los medicamentos relacionados con la enfermedad, así como la polaridad positiva y negativa (representada a través de una iconografía) de cada uno de los fármacos, como se muestra en la Figura 4.2.

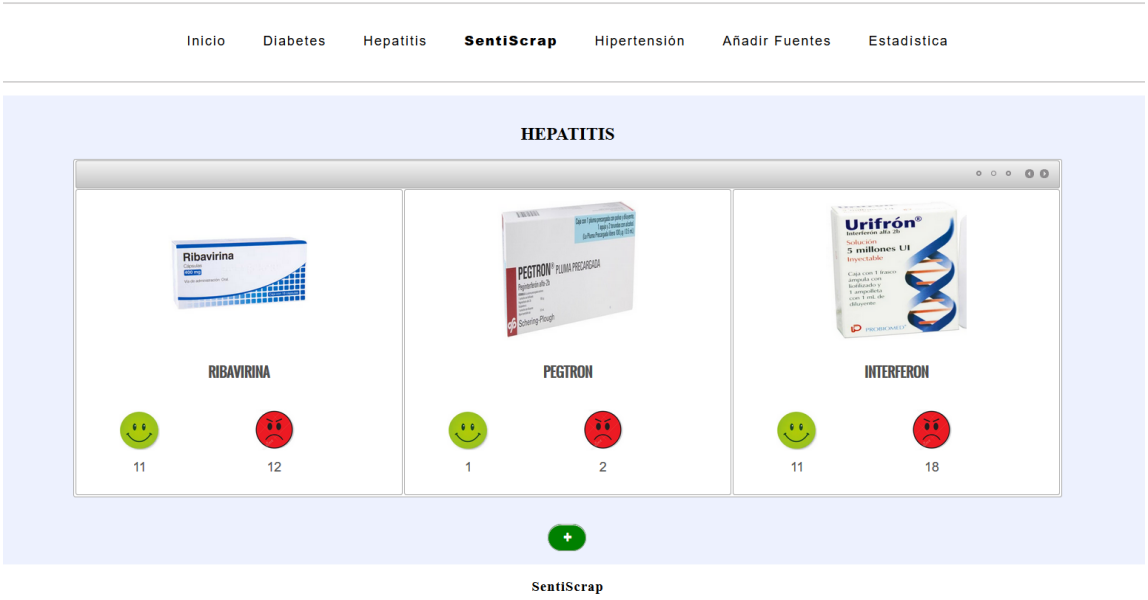

Figura 4. 2 Medicamentos de "Hepatitis"

Para el especialista es necesario conocer los comentarios recolectados, para identificar, evaluar y prevenir los efectos adversos provocados por los medicamentos que toman sus pacientes y así reducir los riesgos a la salud y calidad de vida. Para consultar al repositorio de opiniones da un clic en el botón "+". La respuesta de SentiScrap es proveer todos los comentarios con la polaridad correspondiente, como se muestra en la Figura 4.3.

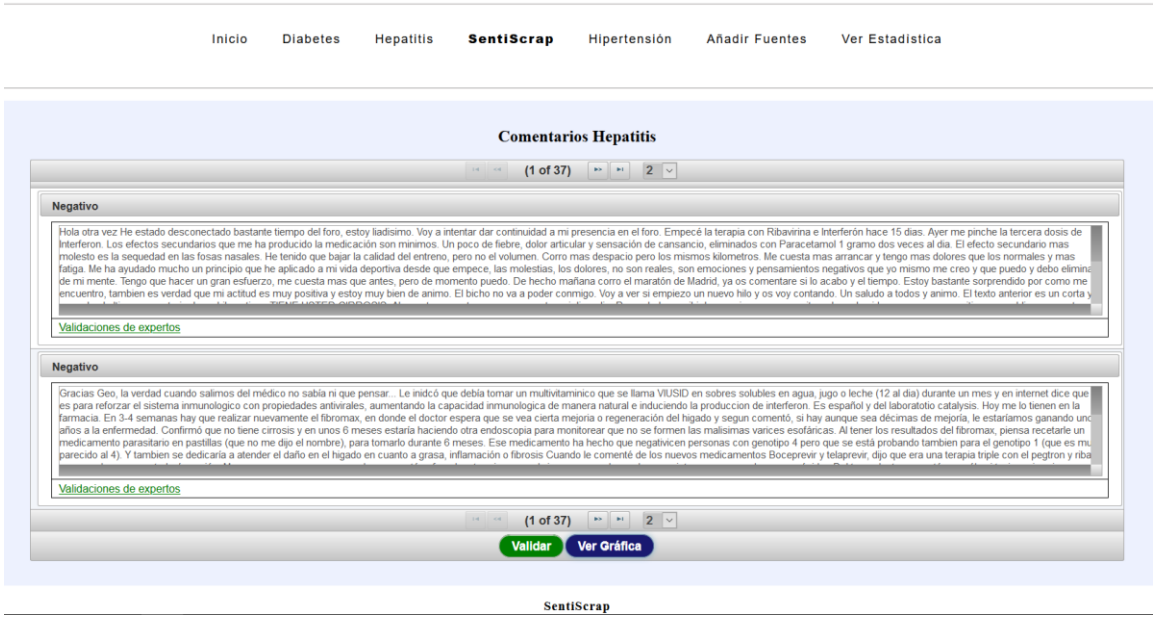

Figura 4. 3 Comentarios de "Hepatitis"

En algunos casos, el especialista en el cuidado de la salud encontrará comentarios mal intencionados, es decir, que alientan la automedicación, que genera consecuencias graves como los efectos adversos, nocivos para la salud. Ante esta situación, se incluyó en este trabajo la validación del experto.

El usuario tiene la posibilidad de validar los comentarios que se muestran en SentiScrap, bajo su experiencia médica, como se muestra en la Figura 4.4, al seleccionar un comentario y dar clic en el botón "Validar".

Figura 4. 4 Validar comentario de "Hepatitis"

El sistema muestra un formulario de datos personales, el cual se llena con los siguientes datos: cédula profesional, nombre y apellidos, para verificar su identidad como se observa en la Figura 4.5.

Figura 4. 5 Formulario de validación para un comentario de "Hepatitis"

La Figura 4.6 muestra el comportamiento del sistema una vez que valida los datos ingresados del especialista, SentiScrap habilita la continuación del formulario, en el cual el experto

selecciona una opción representativa a su opinión sobre el comentario que quiere validar, adicionalmente, será necesario que ingrese un correo electrónico que le permita a los usuarios que leen las validaciones ponerse en contacto ante cualquier duda.

Figura 4. 6 Formulario de validación para un comentario de "Hepatitis" parte 2

Finalmente, el sistema muestra los datos ingresados para que el especialista confirme el envío de los datos al repositorio.

Figura 4. 7 Confirmación de envío de datos

De igual forma, si un experto de la salud quiere colaborar agregando fuentes para su análisis, tiene la opción de agregar un archivo o una sola URL para foros y/o video blogs. Una vez que subió el archivo y este pasó por el proceso de análisis, SentiScrap muestra los siguientes mensajes de éxito, como se observa en la Figura 4.8.

Figura 4. 8 Añadir fuentes para análisis de foros

Por otra parte, cuando el especialista requiere comparar la polaridad entre los comentarios de las enfermedades, el sistema permite visualizar los datos mediante gráficas, que provee la opción "Estadísticas" como muestra la Figura 4.9. En esta vista se presentan gráficos sobre el número de polaridad de los comentarios, positivos y negativos, además de la polaridad correspondiente a cada enfermedad, y el gráfico con las menciones de los medicamentos referidos para cada enfermedad. También se muestra un gráfico circular con los 10 medicamentos más mencionados.

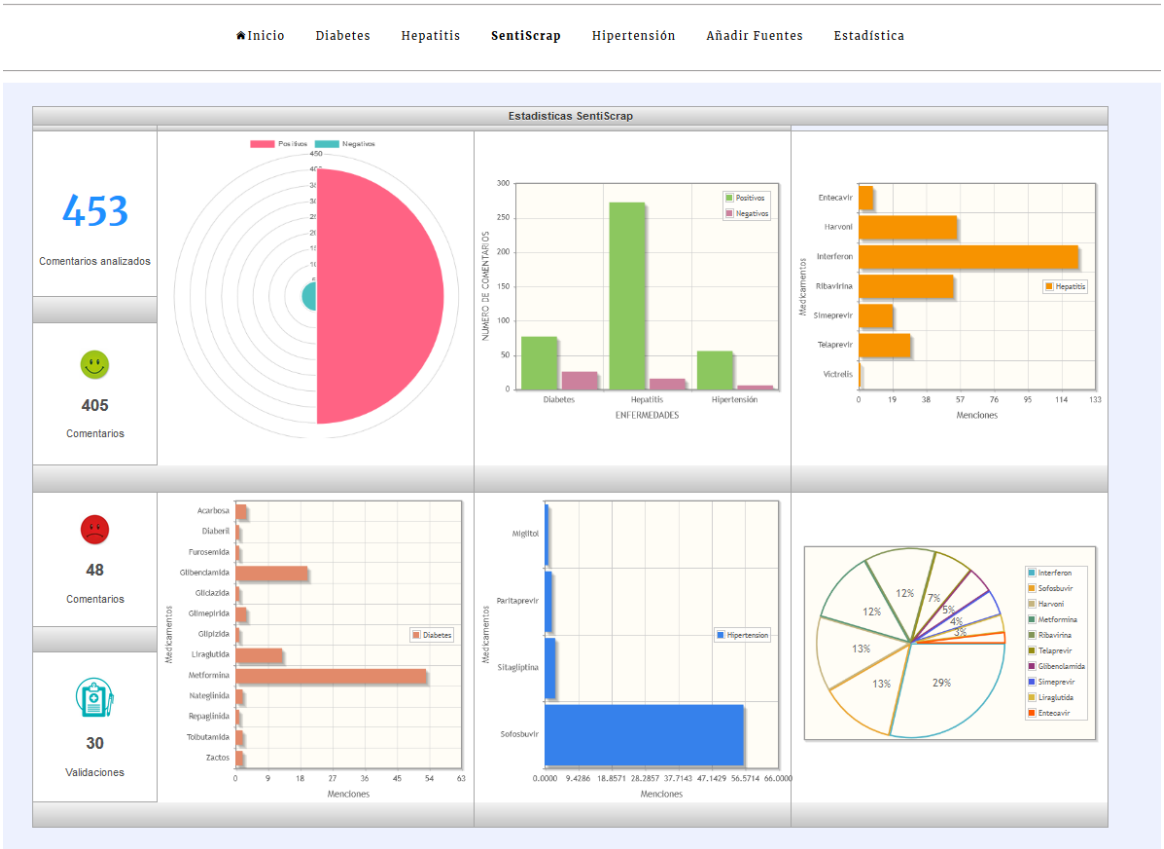

Figura 4. 9 Gráfica de polaridad de las enfermedades

4.1.2 Análisis de polaridad en comentarios publicados en blogs, foros y video blogs en el idioma español de medicamentos para enfermedades crónico-degenerativas bajo el uso de un usuario no experto en el área de la salud.

Cualquier usuario es capaz de acceder a SentiScrap a través de un navegador, ya que se proporciona información tanto para expertos en la salud como para personas que padecen alguna enfermedad, o personas con interés en este tema. El usuario selecciona la enfermedad de su interés dentro del menú que se presenta, como se muestra en la Figura 4.10, en este caso el usuario seleccionó la enfermedad “Diabetes” para ver más detalles.

Figura 4. 10 Inicio de SentiScrap caso usuario

SentiScrap muestra como respuesta a la solicitud del usuario, un carrusel que contienen los medicamentos que se encontraron dentro de los comentarios prescritos para la diabetes con su respectiva polaridad. El usuario, da clic en el botón "+" para conocer los comentarios que hablan acerca de los medicamentos que se le presentan, visible en la Figura 4.11

Figura 4. 11 Medicamentos de "Diabetes"

En la Figura 4.12 se muestra el sistema con los comentarios recolectados a través de foros, blogs y video blogs que se encuentran en la Internet, así como la polaridad correspondiente.

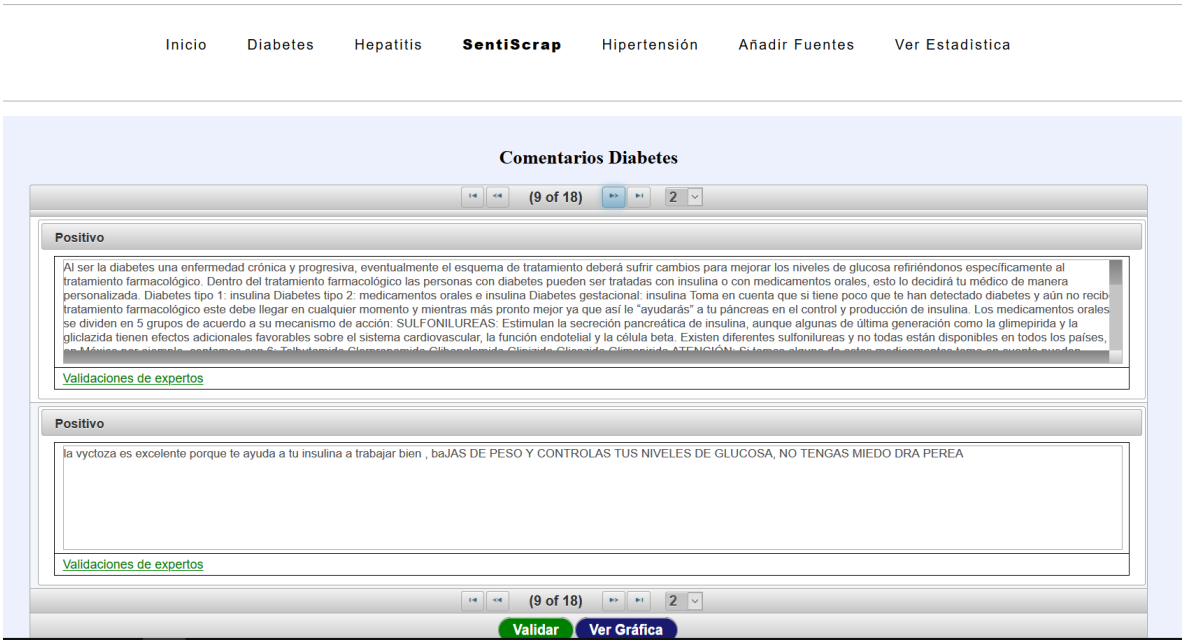

Figura 4. 12 Comentarios de "Diabetes"

Por otra parte, también se le proporcionó al usuario una gráfica que representa la polaridad de la enfermedad seleccionada, como se muestra a continuación en la Figura 4.13.

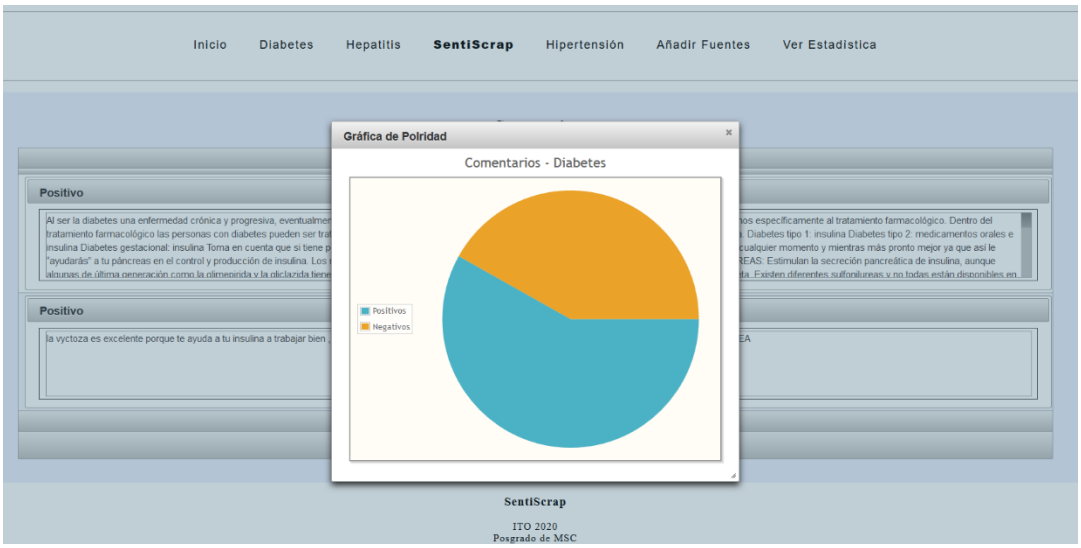

Figura 4. 13 Gráfica de polaridad de "Diabetes"

SentiScrap proporcionó al usuario una opción llamada "Validaciones de expertos" en la que visualiza el nombre y correo electrónico del especialista de la salud que verificó que dichos

comentarios son correctos, o a los que se requiere especial cuidado por ser posiblemente malintencionados y que alientan la automedicación. En la Figura 4.14 se muestra la opción.

Figura 4. 14 Opción para ver "Validaciones de expertos"

También se le facilitó al usuario un botón llamado "Cédulas Profesionales" el cual redirige hacia la página oficial de cédulas profesionales de la SEP (Secretaria de Educación Pública) con el objetivo de facilitar al usuario la corroboración de datos de una fuente oficial, usando los datos presentados en SentiScrap, como se muestra en la Figura 4.15.

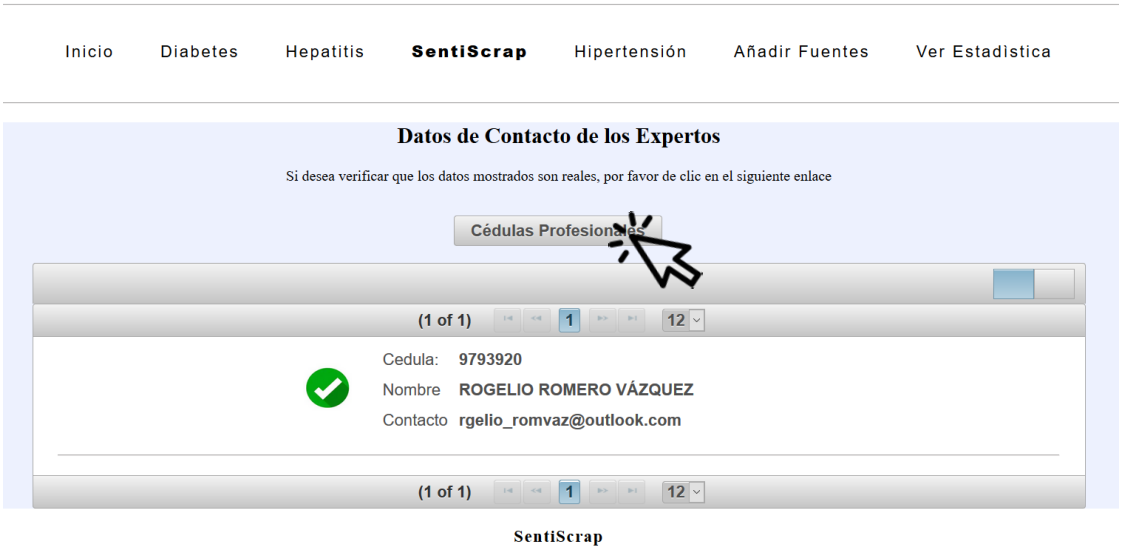

Figura 4. 15 Datos de contacto de los expertos

El sistema permite al usuario colaborar proporcionando fuentes, para incrementar el conocimiento de medicamentos de las enfermedades, así como para la identificación de la polaridad, como se muestra a continuación en la Figura 4.16, en la cual el usuario subió un archivo con las URL de los video blogs que le interesa analizar.

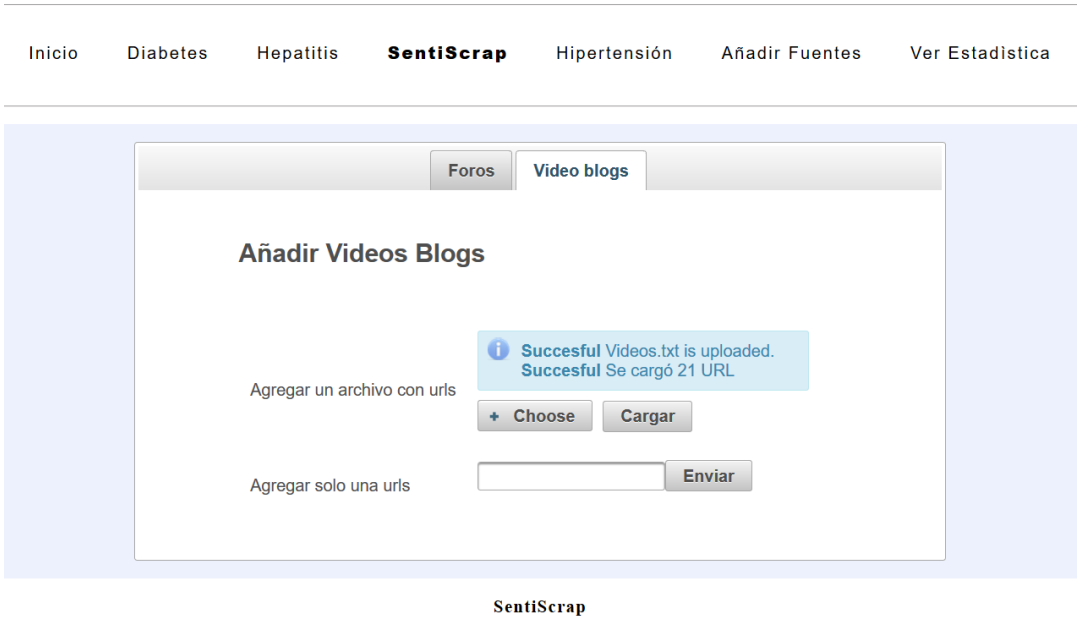

Figura 4. 16 Añadir fuentes de video blogs

Los casos de estudio que se describieron en este capítulo muestran la funcionalidad y la relevancia que tiene SentiScrap como una herramienta que permite conocer la polaridad de los medicamentos para enfermedades crónico-degenerativas, en concreto para la diabetes, hipertensión y hepatitis, además de ser una herramienta activa para la farmacovigilacia cuyo objetivo es la detección, evaluación y prevención de los efectos adversos de la salud para reducir los riesgos a la salud, a través de la validación de los expertos, la cual es sumamente importante para todo aquel usuario que lea los comentarios y quiera conocer una opinión general de un especialista de la salud, sin tener que caer en la automedicación.

4.2 Evaluación y resultados

Para comprobar que la clasificación de la polaridad en los comentarios es favorable, con el enfoque propuesto en esta tesis bajo el sistema denominado SentiScrap, se presenta un análisis de precisión y exhaustividad (*Precision and Recall*).

Esta evaluación consiste en emplear las métricas *Precision and Recall*, que son utilizadas en el dominio de la recuperación de información para medir qué tan bien un sistema informático realiza búsquedas, reconoce patrones y recupera información solicitada por un usuario [52][53].

A continuación en la Tabla 4.1, se presenta una matriz de confusión, que es empleada para resumir el rendimiento de la clasificación del sistema SentiScrap, respecto a las opiniones clasificadas por la polaridad positiva y negativa [54].

Tabla 4. 1 Resultados de la clasificación de opiniones positivas y negativas

	POSITIVOS	**NEGATIVOS**
POSITIVOS	**267**	15
NEGATIVOS	37	**71**
	304	86

4.2.1 Precisión de SentiScrap en la clasificación de las opiniones

La precisión es una medida importante, ya que permite determinar la proporción de identificaciones positivas que son correctas, a través de la siguiente fórmula [55].

$$Precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos\ +\ Falsos\ Positivos} \quad (1)$$

Se calculó la precisión del módulo de clasificación, a partir del conjunto con 390 opiniones, las cuales se revisaron de forma manual, el resultando de la clasificación se muestra en la Tabla 4.1, los datos se usaron para calcular la presión de las opiniones positivas y negativas.

Cálculo de Precisión para las opiniones positivas:

$$Precisión = \frac{267}{267 + 37} = \frac{267}{304} = 0.878 \quad (2)$$

Cálculo de Precisión para las opiniones negativas:

$$Precisión = \frac{71}{71 + 15} = \frac{71}{86} = 0.826 \quad (3)$$

Es decir, el porcentaje de identificación de comentarios positivos es de 87% y de comentarios negativos de un 82%.

4.2.2 Exhaustividad de SentiScrap en la clasificación de las opiniones

Por otra parte, determinar la exhaustividad permite conocer el porcentaje de opiniones que se clasificaron correctamente, se llevó a cabo empleando la siguiente fórmula [55]:

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos\ +\ Falsos\ Negativos} \quad (4)$$

A continuación, se aplica la fórmula *Recall*, utilizando los datos que se muestran en la Tabla 4.1, para opiniones positivas y negativas.

Cálculo de *Recall* para las opiniones con polaridad positiva:

$$Recall = \frac{267}{267 + 15} = \frac{267}{282} = 0.947 \quad (5)$$

Cálculo de *Recall* para las opiniones con polaridad negativa:

$$Recall = \frac{71}{71 + 37} = \frac{71}{108} = 0.657 \quad (6)$$

El porcentaje de opiniones reales que se clasificaron para los comentarios positivos es de un 94% y para los comentarios negativos de un 65%.

4.2.2 Exactitud de SentiScrap en la clasificación de las opiniones

La exactitud o también conocida como *Accuracy* es una métrica que permite la evaluación del modelo usado para la detección de la polaridad de los comentarios recopilados, la exactitud se determinó a través de la siguiente fórmula [56].

$$Accuracy = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones} \tag{7}$$

A continuación, se presenta la fórmula usando los datos presentados en la Tabla 2.1.

$$Accuracy = \frac{267 + 71}{304 + 86} = \frac{338}{390} = \ 0.866 \tag{8}$$

Finalmente, se obtuvo un *Accuracy* en la detección de polaridad de 86%.

Capítulo 5. Conclusiones y recomendaciones

5.1 Conclusiones

A través del análisis del trabajo, presentado en esta tesis, se describió la importancia que tiene SentiScrap, dado que hay una gran oportunidad de aplicación de análisis de opiniones y detección de polaridad en el idioma español, especialmente en sitios Web especializados, blogs y video blogs dentro del contexto del cuidado de la salud, en especial de la farmacovigilancia, ya que son una fuente de obtención de datos que no se ha explorado a profundidad.

El propósito de este trabajo es analizar las técnicas y herramientas necesarias para realizar el *Scraping* a blogs, video blogs y sitios web especializados, además, para la detección de polaridad de las opiniones recopiladas. Dicho análisis fue indispensable para realizar posteriormente una arquitectura *ad hoc* con la propuesta de tesis, que cumpliera con todos los objetivos planteados.

La arquitectura planteada permitió el correcto desarrollo de la aplicación web SentiScrap, la cual tiene como propósito proveer información a los especialistas de la salud, acerca de la polaridad y de las opiniones encontradas en la Web, sobre enfermedades crónico-degenerativas, en concreto, hepatitis, diabetes e hipertensión, con dicha información, el experto identifica lo que las personas opinan sobre los medicamentos prescritos para las enfermedades mencionadas y con ello, dar una validación (ingresando sus nombre completo y cédula) sobre cada opinión que muestra SentiScrap, con el objetivo primordial de hacer notar cuando un comentario es mal intencionado y que alienta la automedicación o cuando el comentario es correcto y no representa riesgo a la salud. Por otra parte, SentiScrap muestra a través de un icono la validación del experto y sus datos para que los usuarios no expertos en el área de la salud puedan ponerse en contacto directo con un médico ante alguna duda y así evitar la automedicación y el autodiagnóstico.

Cabe destacar que es importante incrementar el conocimiento que presenta SentiScrap, lo cual es posible para los usuarios, los cuales agregan una URL o archivos con las URL de los blogs y video blogs para ser analizados, aunque es una tarea que resulta ser tardada, por la obtención de las opiniones, el tratamiento y el análisis que conlleva agregar nuevas fuentes, los experimentos muestran resultados alentadores, ya que se obtuvo una *Precision* del 87% y 82% para opiniones

positivas y negativas, respectivamente, y un *Recall* del 94% para opiniones positivas y 65% para opiniones negativas, además, el *Accuracy* del modelo usado para la detección de polaridad de un 86%.

5.2 Recomendaciones

Aunque el trabajo de tesis se realizó con éxito, en tiempo y forma, cumpliendo con cada uno de los objetivos planteados, se identificaron algunas mejoras que con su implementación harán de SentiScrap una aplicación más completa, a continuación, se describen dichas mejoras.

Como recomendación, es viable incluir más aspectos a analizar, como la obtención de los efectos adversos que se mencionen en los comentarios, con el objetivo de proveer dicha información a la comunidad médica. Además, incluir el análisis del precio de los medicamentos, así como la identificación del tiempo de tratamiento que toman los pacientes que comentan en foros, blogs, y video blogs sobre las enfermedades crónico-degenerativas abordadas.

Por otra parte, agregar nuevas fuentes de datos a analizar, como lo son las redes sociales, es adecuado para tener un sistema enriquecido de comentarios en el idioma español y de dominio médico que permita a los especialistas de la salud conocer la opinión de la comunidad en relación a los medicamentos prescritos.

Finalmente, se recomienda agregar más enfermedades crónico-degenerativas a analizar como el cáncer y el asma, por mencionar algunos, y en efecto incrementar los nombres de medicamentos en la ontología para realizar la identificación de dominio.

ANEXOS

Gestión de Componentes de Software

Asunto: Carta de uso de sistema
Orizaba, Ver. a 5 de julio del 2020

Dr. José Luis Sánchez Cervantes
Catedrático CONACYT
Instituto Tecnológico de Orizaba

Por medio de la presente me permito enviar a usted un cordial saludo, así mismo me permito informarle del interés y compromiso de la empresa GECOSOFT S.A. de .C.V por brindar el apoyo institucional necesario para llevar a cabo el proyecto "*Sistema de Análisis de Opiniones en Español de medicamentos para enfermedades crónico-degenerativas basado en técnicas de Web Scraping*".

El desarrollo de este proyecto es llevado a cabo por la *ISC. Karina Castro Pérez* como parte de su tesis de Maestría en Sistemas Computacionales, así como por investigadores del Instituto Tecnológico de Orizaba en vinculación con GECOSOFT S.A. de C.V.

Esperando obtener los beneficios que emanen de la colaboración entre el Instituto Tecnológico de Orizaba y GECOSOFT S.A. de C.V. me despido de usted quedando a sus finas atenciones.

ATENTAMENTE

Mtro. Jorge Ernesto González Díaz
Director de Proyectos de Innovación y
Desarrollo Tecnológico
GECOSOFT S. A DE C.V.

GECOSOFT, S.A. DE C.V.
RFC: GEC-121218-EY4
SUR 2 No.225 - INT 3, COL. CENTRO,
ORIZABA, VER.
TEL. (272) 147-89-00

www.gecosoft.com.mx

Sur 2 No. 225-3 Orizaba, Veracruz Teléfono y Fax (272) 726 44 80

Referencias

[1] B. Baesens and S. Broucke, “Web Scraping Basics,” in *Practical Web Scraping for Data Science*, New York: Apress, 2018, pp. 3–4.

[2] D. Neef, “The Data Collectors,” in *Digital Exhaust*, 1st ed., New Jersey: Pearson, 2015, p. 151.

[3] V. Smith, “Introducing Web Scraping and Go,” in *Go Web Scraping Quick Start Guide*, 1st ed., U.K: Packt, 2019, pp. 5–6.

[4] B. Liu, “Web Data Mining,” in *Web Data Mining*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 311.

[5] B. Baesens and S. vanden Broucke, “From Web Scraping to Web Crawling,” in *Practical Web Scraping for Data Science*, 1st ed., Apress, 2018, pp. 155–156.

[6] J. Cho and H. Garcia-Molina, “Parallel crawlers,” in *Proceedings of the 11th International Conference on World Wide Web*, Honolulu, Hawaii, USA, 2002, pp. 124–135, doi: 10.1145/511446.511464.

[7] L. Y. Zhang, T. Fei, X. Zhang, Y. Li, and J. Yi, “Communications and Information Processing,” in *Communications in Computer and Information Science*, vol. 288, M. Zhao and J. Sha, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 430–437.

[8] N. Yadav, S. K. Divyaa, and S. K. Dubey, “Evaluation of Faculty Performance in Education System Using Classification Technique in Opinion Mining Based on GPU,” in *Computational Intelligence in Data Mining—Volume 2*, 2015, vol. 2, no. December, pp. 109–111, doi: 10.1007/978-81-322-2731-1.

[9] B. Liu, “Sentiment Analysis: A Fascinating Problem,” in *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012, pp. 7–15.

[10] B. Marr, “A = Apply Analytics,” in *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*, U.K: John Wiley & Sons,

2015, p. 113.

[11] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 3934–3942, 2013, doi: 10.1016/j.eswa.2012.12.084.

[12] M. Thangavel and P. Karthikeyan, "Ontology Based Feature Extraction from Text Documents," in *Applications of Security, Mobile, Analytic, and Cloud (SMAC) Technologies for Effective Information Processing and Management*, Pensilvania: IGI Global, 2018, p. 175.

[13] M. Bajee and S. Zitnik, "Text Mining in Medicine," in *Computational Medicine in Data Mining and Modeling*, New York: Springer, 2013, pp. 105–106.

[14] "OWL Web Ontology Language Overview." https://www.w3.org/TR/2004/REC-owl-features-20040210/#s1.2 (accessed Mar. 14, 2019).

[15] "OWL 2 Web Ontology Language Primer (Second Edition)." https://www.w3.org/TR/owl2-primer/#What_is_OWL_2.3F (accessed Mar. 26, 2019).

[16] J. Thanaki, "Feature Engineering and NLP Algorithms," in *Python Natural Language Processing*, Packt, Ed. U.K, 2017, pp. 102–172.

[17] A. Bolshoy, Z. (Vladimir) Volkovich, V. Kirzhner, and Z. Barzily, "Mathematical Models for the Analysis of Natural-Language Documents," in *Genome Clustering: From Linguistic Models to Classification of Genetic Texts*, A. Bolshoy, Z. (Vladimir) Volkovich, V. Kirzhner, and Z. Barzily, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 25–26.

[18] M. D. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. Á. Rodríguez-García, and R. Valencia-García, "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach," *Comput. Math. Methods Med.*, vol. 2017, 2017, doi: 10.1155/2017/5140631.

[19] J. Matoušek, Z. Hanzlíček, M. Campr, Z. Krňoul, P. Campr, and M. Grůber, "Web-Based System for Automatic Reading of Technical Documents for Vision Impaired Students," 2011, vol. 6836, p. 369, doi: 10.1007/978-3-642-23538-2_46.

[20] J. Benesty, M. M. Sondhi, and Y. A. Huang, "Linguistic Processing for Speech Synthesis," in *Springer Handbook of Speech Processing*, Springer, Ed. 2007, p. 459.

[21] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, 2005, pp. 573–580, doi: 10.3115/1219840.1219911.

[22] J. Torruella and J. Llisterri, "Diseño de corpus textuales y orales," *Filol. e Informática. Nuevas Tecnol. en los Estud. filológicos*, pp. 45–77, 1999, [Online]. Available: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf.

[23] H. Spona, "Bases de datos y MySQL Fundamentos," in *Programación de bases de datos con MySQL y PHP*, Marcombo, 2010, pp. 81–82.

[24] S. A. Carrillo, N. M. Ruiz, J. M. M. Rodriguez, O. P. Capote, and A. V. Miranda, "Introducción a las Bases De Datos," in *Introducción a las bases de datos: el modelo relacional*, 1st ed., España: Paraninfo, 2005, pp. 6–7.

[25] M. V. N. Cabello, "Las Bases de Datos," in *Introduccion a las Bases de Datos relacionales*, España: Vision Libros, 2010, pp. 34–35.

[26] A. Cobo, "Sistemas de Bases De Datos. Gestores De Bases de Datos (SGBD). Arquitecturas," in *Diseño y programación de bases de datos*, España: Vision Libros, 2007, pp. 7–8.

[27] M. Q. Ugalde, "Los Foros Como Recurso para la Discusión y el Interccambio," in *Nuevas tecnologías: Procedimientos básicos e ideas de aplicación en educación especual*, Costa Rica: EUNED, 2004, pp. 100–101.

[28] E. Montero, L. Rodríguez-Carmona, and Á. Herrero, "Recursos audiovisuales en la

docencia a nivel universitaria. El uso del viodeblog como herramienta de comunicación," *Hist. y Comun.*, vol. 18, p. 166, 2013.

[29] K. Lee *et al.*, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," *26th Int. World Wide Web Conf. WWW 2017*, pp. 705–714, 2017, doi: 10.1145/3038912.3052671.

[30] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," *2017 1st Int. Work. Arab. Scr. Anal. Recognit.*, pp. 114–118, Feb. 2017, doi: 10.1109/asar.2017.8067771.

[31] D. Cavalcanti, "Progress in Artificial Intelligence," vol. 10423, pp. 815–827, 2017, doi: 10.1007/978-3-319-65340-2.

[32] V. Gopalakrishnan and C. Ramaswamy, "Patient opinion mining to analyze drugs satisfaction using supervised learning," *J. Appl. Res. Technol.*, vol. 15, no. 4, pp. 311–319, 2017, doi: 10.1016/j.jart.2017.02.005.

[33] Y. Peng, M. Moh, and T. S. Moh, "Efficient adverse drug event extraction using Twitter sentiment analysis," *Proc. 2016 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2016*, pp. 1011–1018, 2016, doi: 10.1109/ASONAM.2016.7752365.

[34] S. Sabra, K. Mahmood Malik, and M. Alobaidi, "Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives," *Comput. Biol. Med.*, vol. 94, pp. 1–10, 2018, doi: 10.1016/j.compbiomed.2017.12.026.

[35] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artif. Intell. Med.*, vol. 64, no. 1, pp. 17–27, 2015, doi: 10.1016/j.artmed.2015.03.006.

[36] C. Wu, J. Liu, F. Wu, Y. Huang, Z. Yuan, and X. Xie, "MSA: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention," *WSDM 2019 - Proc. 12th ACM Int. Conf. Web Search Data Min.*, pp. 33–41, 2019, doi: 10.1145/3289600.3290980.

[37] S. Karimi, A. Metke-Jimenez, and A. Nguyen, “CADEminer: A system for mining consumer reports on adverse drug side effects,” in *ESAIR 2015 - Proceedings of the 2015 Workshop on Exploiting Semantic Annotations in Information Retrieval, co-located with: CIKM 2015*, Melbourne, Australia, 2015, pp. 47–50, doi: 10.1145/2810133.2810143.

[38] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, “Sentiment analysis and the complex natural language,” *Complex Adapt. Syst. Model.*, vol. 4, no. 1, 2016, doi: 10.1186/s40294-016-0016-9.

[39] M. L. Cabling *et al.*, “Sentiment Analysis of an Online Breast Cancer Support Group: Communicating about Tamoxifen,” *Health Commun.*, vol. 33, no. 9, pp. 1158–1165, 2018, doi: 10.1080/10410236.2017.1339370.

[40] Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah, and A. Shah, “Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis,” *2018 IEEE 5th Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2018*, pp. 1–4, 2019, doi: 10.1109/ICETAS.2018.8629198.

[41] H. Luna-Aveiga *et al.*, “Sentiment polarity detection in social networks: An approach for asthma disease management,” in *Advances in Intelligent Systems and Computing*, 2018, vol. 629, pp. 141–152, doi: 10.1007/978-3-319-61911-8_13.

[42] S. Rathi, S. Shekhar, and D. K. Sharma, “Opinion mining classification based on extension of opinion mining phrases,” in *Advances in Intelligent Systems and Computing*, 2016, vol. 408, pp. 717–724, doi: 10.1007/978-981-10-0129-1_74.

[43] J. Hedley, “jsoup: Java HTML Parser,” 2009. https://jsoup.org/ (accessed Jun. 25, 2020).

[44] “JDOM: Misión.” http://www.jdom.org/mission/index.html (accessed Oct. 19, 2019).

[45] “Stanford CoreNLP - Software de lenguaje natural | Stanford CoreNLP.” https://stanfordnlp.github.io/CoreNLP/index.html#about (accessed May 01, 2019).

[46] “protégé.” https://protege.stanford.edu/products.php (accessed May 01, 2019).

[47] “Apache Jena - An Introduction to RDF and the Jena RDF API.” https://jena.apache.org/tutorials/rdf_api.html (accessed Oct. 19, 2019).

[48] “PostgreSQL: Acerca de.” https://www.postgresql.org/ (accessed Jun. 25, 2020).

[49] T. Groussard, “Presentación,” in *JAVA 7: Los fundamentos del lenguaje Java*, Barcelona: Editions ENI, 2012, pp. 9-10–11.

[50] “NetBeans IDE - Descripción general,” 2018. https://netbeans.org/features/index.html (accessed Jun. 25, 2020).

[51] C. Nieves-Guerrero, J. Ucán-Pech, and V. Menéndez-Domínguez, “UWE en Sistema de Recomendación de Objetos de Aprendizaje. Aplicando Ingeniería Web: Un Método en Caso de Estudio,” *Rev. Latinoam. Ing. Softw.*, vol. 2, no. 3, p. 137, Apr. 2014, doi: 10.18294/relais.2014.137-143.

[52] K. M. Ting, “Precision and Recall,” in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, p. 781.

[53] G. I. W. Claude Sammut, “Precision,” in *Encyclopedia of Machine Learning and Data Mining*, Second Edi., Springer, Boston, MA, 2017, p. 781.

[54] K. M. Ting, “Confusion Matrix,” in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, p. 209.

[55] B. Carterette, “Precision and Recall,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 2126–2127.

[56] G. I. W. Claude Sammut, “Accuracy,” in *Encyclopedia of Machine Learning*, p. 9.