EXTRACCIÓN DE SIGNOS Y SÍNTOMAS DE ENFERMEDAD MEDIANTE CONSULTA WEB

EXTRACTION OF SIGNS AND SYMPTOMS OF DISEASE THROUGH WEB CONSULTATION

Pérez Escamilla Javier[a], Cuaya Simbro German[b], Cruz Guerrero René[b], Mendoza Guzmán Lorena[a]

[a] Tecnológico Nacional de México/ITSOEH, División de Ingeniería en Sistemas Computacionales. Mixquiahuala de Juárez, Hidalgo, México. C.P.42700.
[b]Tecnológico Nacional de México/Instituto Tecnológico Superior del Oriente del Estado de Hidalgo, División de Ingeniería en Sistemas Computacionales., Apan, Hidalgo, México. C.P. 43900.
javierperez@itsoeh.edu.mx

RESUMEN. *En tiempos recientes debido a la condición de pandemia que vive la humanidad, la identificación de síntomas asociados al padecimiento de COVID-19 se ha convertido en un tema de suma importancia. El presente trabajo propone un proceso que permite extraer los síntomas de esta enfermedad a partir de la información disponible en publicaciones WEB, realizando las adecuaciones y transformaciones necesarias para contar con elementos que permitan conocer, de manera simple y dinámica, la asociación que existe con un padecimiento. La propuesta, es una herramienta de apoyo a personal médico, que ha sido probada identificando sintomatología del COVID-19. En esta, se ha detectado como factor común la gran diversidad de información en publicaciones no científicas, destacando que en ellas hay un gran número de observaciones realizadas por médicos y pacientes. El trabajo aborda la dificultad de encontrar nuevos síntomas mediante: el uso de agentes, para la extracción de información de sitios de interés; una clasificación de textos usando Máquinas de Vector de Soporte , para la identificación de párrafos relevantes; Procesamiento de Lenguaje Natural en la extracción de términos, usando una función de etiquetado e identificación de sustantivos; expresiones regulares que facilitan el reconocimiento de entidades, mediante el uso de frecuencias de ítems; soporte de conocimiento adquirido, en una colección de palabras de sintomatología; y la presentación de resultados en forma visual. El modelo alcanza un 82% y 80% en precisión y recuerdo en la clasificación y un diccionario de datos de 704 entradas, consultado a expertos en salud. La información obtenida se muestra en una imagen, esto ayuda al personal médico a revisar o identificar cualquier nueva evidencia., resaltando los de menor incidencia o nuevos del padecimiento.*

Palabras clave: *Consulta Web, Nube de Palabras, Signo y síntoma*

ABSTRACT. *In recent times, due to the pandemic condition that humanity is experiencing, the identification of symptoms associated with the suffering of COVID-19 has become a matter of utmost importance. The present work proposes a process that allows to extract the symptoms of this disease from the information available in WEB publications, making the necessary adjustments and transformations to have elements that allow knowing, in a simple and dynamic way, the association that exists with the disease. The proposal is a support tool for medical personnel, which has been tested by identifying symptoms of COVID-19. In this, the great diversity of information in non-scientific publications has been detected as a common factor, highlighting that there are a large number of observations made by doctors and patients. The work addresses the difficulty of finding new symptoms through: the use of agents to extract information from sites of interest; a text classification using Support Vector Machines, for the identification of relevant paragraphs; Natural Language Processing in the extraction of terms, using a function of labeling and identification of nouns; regular expressions that facilitate the recognition of entities, through the use of item frequencies; support of acquired knowledge, in a collection of symptomatology words; and the presentation of results in visual form. The model achieves 82% and 80% accuracy and recall in classification and a 704-entry data dictionary, consulted with health experts. The information obtained is shown in an image, it helps the medical personnel to review or identify any new evidence, highlighting those with a lower incidence or new ones of the disease.*

Key words*: Web Query, Wordcloud, Sign and Symptom.*

INTRODUCCIÓN

En México, no existen datos científicos que antecedan a investigaciones relacionadas a la detección de sintomatología mediante búsquedas en internet. Investigaciones en otros países, han permitido tener un mejor panorama sobre este tema, tomando en cuenta la diversidad de síntomas y la forma en la que se recaban los datos[1]. Dichos estudios presentan una deficiencia generalizada en cuanto a las técnicas de recolección de datos, debido a que estos se obtienen vía centros médicos o vía encuesta[2, 3], Esta información, puede generar una disociación, incomodidad y angustia en las personas que presentan algún cuadro clínico, dado que el entendimiento propio de un padecimiento, puede asociarse a otro, o en su caso causar complicaciones graves e incluso la muerte[4]. Así entonces, toma importancia la falta de conocimiento médico sobre la enfermedad. Además, la propia sintomatología, puede no presentarse de igual forma en una persona que en otra[5].

Una vez analizada esa situación, se optó por utilizar una extracción web de información de un padecimiento, en este caso COVID-19, aplicando una técnica de clasificación, Procesamiento de Lenguaje Natural y agentes. Sustentado, en que las personas recurren con frecuencia a las noticias en la red, para visualizar y externar acontecimientos de la vida cotidiana, incluso aquellos relacionados con padecimientos graves[6]. Cada individuo realiza su interpretación de la información, además, si es bajo formato de entrevista o similar, puede ser reinterpretada, pero sus síntomas tendrán relación o servirán de información a la audiencia lectora.

Importancia

No hay una herramienta automática de uso libre en México, que permita revisar e identificar síntomas de un padecimiento de interés. Existen diferentes textos publicados en la red, pero algunos de los usuarios de internet, no consultan fuentes confiables. Además, no en todos los casos es posible procesar toda la información que se tiene disponible.

Surgen nuevas enfermedades en las que no hay una patología definitiva, ni se identifican todos los riesgos asociados, lo que puede llevar a una situación pandémica[7]. Los signos y síntomas pueden ser desapercibidos. Entonces, para mitigar la mortandad y efectos adversos, son necesarios conocimientos para el personal médico y de apoyo sustentados en un sistema inteligente[8]. Por lo tanto, información visual y de fácil entendimiento, apoya en que se concientice a tomar acciones de prevención.

Problemática

La herramienta propuesta, está diseñada para servir de apoyo al sector salud y al público en general.

Clasificar textos orientados de una patología en común, acompañado de un diccionario de datos y un agente para la extracción de información de sitios de interés, permitirá tener información vigente y de nuevos datos. Procesar la información, mediante técnicas de lenguaje natural para la identificación de entidades, permite que se pueda desplegar una nube de palabras que contenga los aspectos relevantes o nuevos. Así, un usuario podrá visualizar la información y tomar acciones de contención, pudiendo prevenir síntomas graves, secuelas e incluso la muerte.

La metodología para la interpretación de síntomas de un padecimiento se realizó en dos etapas. La primera etapa consiste en la recolección de fuentes y textos asociados en publicaciones que contengan síntomas. Allí, se extrae el signo y síntoma, se acumula en una base de conocimiento. En la segunda etapa se realiza un proceso de clasificación para la extracción de textos, un proceso de toquenización, selección de palabras y publicación de resultados.

MARCO TEÓRICO

A continuación, resaltaremos elementos usados en la investigación. Enlistamos aquellos conceptos y técnicas aplicados que sustentan el método.

A) Inteligencia artificial

AI, por sus siglas en inglés de *Artificial Intelligence*, refiere a una máquina que puede imitar una acción humana, donde se maximicen las posibilidades de éxito, realizado de manera autónoma y aplicando el conocimiento[9]. Siempre que una máquina realice acciones humanas, se considera un elemento inteligente, un autómata.

B) Aprendizaje de máquina

ML, por sus siglas en inglés de *Machine Learning*. Es un grupo de algoritmos, dentro del IA, que realizan tareas automáticas sin la intervención explícita de un humano[10]. Se definen varias categorías: algoritmos supervisados, que incluyen la clasificación y regresión; los algoritmos no supervisados, como el agrupamiento, asociación y reducción de dimensión.

Dentro del ML, Máquinas Vector de Soporte: SVM, de sus siglas en ingles de *Support Vectorial Machine*, es un algoritmo de ML para clasificación. Transforma la entrada en un espacio n-dimensional, para ello se especifica una función de *Kernel*, así entonces, identifica las características transformadas en dos instancias de clase, el número de características está determinando por el número de vectores de soporte[11].

SVM, se puede sustentar en RBF, de sus siglas en ingles de *Radial Base Function.* En Ec. 1 define la función, donde σ representa la anchura del *Kernel*. Así entonces podrá definirse un espacio dimensional en la clasificación binaria, donde toda "*x*" está representada por todas las instancias $x_1, \ldots, x_n$[12].

$$(RBF) = K(x_1, x_2) = \exp\left(-\frac{||x_1 - x_2||^2}{2\sigma^2}\right) \quad (1)$$

Clasificación binaria es una técnica de ML, se basa en un conjunto que contiene dos clases: pertenencia o no pertenencia. Así, una nueva instancia que se presume del conjunto asociado, podrá asignarsele una etiqueta (0,1)[13]. Múltiples técnicas son aplicadas para la clasificación binaria, varias métricas son asociadas en el proceso.

Métricas de evaluación: en la clasificación, refiere a las instancias correctamente clasificadas y no correctamente clasificadas. TP o Verdaderos positivos, FP falsos positivos, TN o Verdaderos Negativos, FN o falsos negativos, las métricas asociadas son *Accuracy, Recall y Precision*[14]. Asociados con la exactitud, el recuerdo y la precisión:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Presicion = \frac{TP}{TP+FP} \quad (4)$$

- Ec. 2, mide la exactitud.
- Ec. 3, mide el recuerdo.
- Ec. 4, mide precisión del modelo.

Dónde: TP, son las instancias verdaderas clasificadas correctamente; TN, son las instancias verdaderas clasificadas incorrectamente; FP, son las instancias falsas clasificadas correctamente; FN, son las instancias falsas clasificadas incorrectamente.

Dentro de las métricas, Ec. 1 representa la exactitud en la clasificación, es decir lo acertado del modelo, Ec. 2 representa la capacidad del modelo de discriminar entre las clases, Ec. 3 representa, la predicción correcta de instancias.

C) Agentes

Los agentes son parte de la automatización de tareas, el fundamento es un Agente Racional, un elemento capaz de razonar. Así entonces, una Agente Inteligente es un algoritmo capaz de percibir su entorno, tiene sensores, actuadores y es tiene un método de razonamiento[15]. El sensor, puede ser un método o un dispositivo, interactúa con el medio, el actuador, es un método o dispositivo que realiza acción. El método de razonamiento contiene un algoritmo que es capaz de proyectar la percepción dado el entorno y prepara las acciones que ejecuta el actuador.

D) Raspado web

Dada la gran cantidad de páginas web disponibles en la red, se requieren de conocimiento de muchas estructuras para interpretar la información valiosa o de interés. Es por ello que el raspado web o *Web Scrapping*, es una herramienta que facilita la extracción de información HTML, *Hiper Text Markup Language*. Así entonces, puede acceder a una dirección en internet y obtener meta información para el procesamiento posterior[16].

E) Wordcloud

Wordcloud es una técnica para la visualización y la representación gráfica de información, en este caso, palabras. La frecuencia de palabras y una estructura de diccionario, representan la entrada para ésta

técnica. Allí, las palabras tomarán relevancia en función de los textos analizados[17].

F) Procesamiento del Lenguaje Natural
NLP, por sus siglas en inglés de *Natural Language Processing*. Es una de las ramas de la AI que se encarga de procesamiento de la lingüística y su aplicación en las computadoras, usando técnicas para el análisis y comprensión del lenguaje natural[18].

G) Conjunto de entrenamiento y prueba
Conjunto de entrenamiento: para la clasificación de texto, se requiere un conjunto de tuplas, que deberán contener la característica de lo que se desea clasificar, además contendrán un atributo que será la etiqueta de clase de pertenencia. Conjunto prueba: son tuplas que sirven como referencia para probar la validez del modelo, Generalmente corresponde a una parte del conjunto de entrenamiento[19].

En el procesamiento de conjuntos y técnicas de ML, existen librerías como Sklearn. Esta última, es un conjunto de algoritmos y funciones para la carga de datos, procesamiento, tareas de clasificación y métricas de evaluación. La aplicación de ella, se extiende a más actividades en AI[20].

H) Análisis sintáctico, tokenización, lematización y raíces de palabra
El análisis sintáctico, refiere al estudio de las palabras, para así obtener una interpretación de la oración, así mismo la concordancia y jerarquía[21].

Tokenización, es la normalización de textos, donde se realiza una segmentación de las palabras, tomando en cuenta aquellos elementos de interés para el estudio que se pretenda realizar[22].

Lematización, es la técnica que permite extraer la forma canónica de una palabra, así entonces se podrá relacionar la palabra derivada. Raíces de palabra o *stemming*, es una técnica para obtener la raíz de la palabra[23]. Obtener la raíz de la palabra nos ayuda a darle significado a la oración.

I) NLTK, corpus y Average Perceptron Tagger
NLTK, por sus siglas en inglés de *Natural Language Toolkit*, es una librería para el procesamiento de texto construida en el lenguaje *Python* para la clasificación, tokenización, derivación, etiquetado, análisis sintáctico y de razonamiento sintáctico[24].

Corpus, es un gran cuerpo de textos, contiene un conjunto de palabras en un lenguaje semi-estructurado. Aquí se aplica el análisis estadístico y pruebas de hipótesis, donde se validan las reglas lingüísticas[25].

APT, de sus siglas en inglés de *Average Perceptron Tagger*. Es una función de NLTK para el etiquetado de las palabras en la gramática tradicional como lo son sustantivos, verbos, adjetivos, pronombres, etc. se requiere aplicar cuando se necesita identificar una parte de la oración que se está tratando. Así entonces, se obtiene información gramatical y de significado para su posterior procesamiento[26].

TRABAJOS RELACIONADOS

A continuación, revisaremos trabajos previos, resaltamos aportes que se han realizado en Latinoamérica.

López (2020) identifico las sintomatologías generales, al trabajar en la identificación de reportes médicos, los cuales procesó y generó una base de conocimiento llamada *Addese*, que es una extensión de Base de Datos Sintácticos del español actual (BDS). Aplicó una estrategia de corpus para lograr la clasificación, identificando síntoma y signo. Así entonces, usó técnicas léxico combinatorias para generar un nuevo conjunto de datos, basando en conocimiento experto, sin embargo, sólo lista los elementos que están dentro del contexto, no identifica nuevos síntomas, ni lo enfoca a una patología.[27]

Sánchez y Velasco, (2014) implementan un proyecto llamado *Varimed*, usando como base el conocimiento experto, y un análisis de texto. Identifica, mediante las partes de la oración, el contexto al que refiere la palabra, la significancia de la misma y la localización del padecimiento. Afronta el proceso de crear *Corpus* y acaba generando una taxonomía de clasificación. Sin embargo, no abona al lenguaje del español y no aborda al contexto de la implementación de una nueva detección automática de síntomas asociados[28].

Liu et al, (2020) Aborda la tarea de la detección temprana de choques sépticos. Aplica un modelo de regresión y técnicas de NPL, tomando como referencia, una base de datos que contiene información de sensores médicos y estados psicológicos de estos pacientes. Se da a la tarea de pre procesar la información, obteniendo alta predicción en la condición de pre choque séptico. No aborda la tarea de identificar el padecimiento o síntomas[29].

Hussan, Choi y Lee, (2020) Proponen un modelo para la detección de padecimiento. Aplicando técnicas de *NPL* para la extracción de tokens y bigramas, procesa los datos en un diccionario medico de acceso restringido llamado UMLS, *Unified Medical Language System*, así es capaz de listar un conjunto de padecimientos asociados a síntomas. Deja de lado la correcta identificación y usa un proceso de negación léxica para reducir el tamaño de la salida[30].

Hass et al, (2020) abordan la tarea de identificar síntomas asociados a la psicosis, toma como base elementos lingüísticos extraídos con técnicas de NPL, información clínica y de neuro-imágenes. Aplica el método *SCCA*, del inglés *Sparse Canonical Correlation Analyses*, demostrando la correlación en los elementos lingüísticos en la coherencia y conductas dirigidas a objetivos con la psicosis. No aborda el problema de identificar nuevos síntomas, centrando su trabajo en explorar las relaciones del habla con las conexiones neuronales[31].

Cruz, Maña y Mata, (2010) Revisan las capacidades de las expresiones regulares versus la técnicas de negación de NPL para identificar negaciones y especulaciones. Aborda la tarea de la clasificación de textos médicos de *BioScope*. Tomando como referencia la base de medida F_1, demuestra que el uso de expresiones regulares y un diccionario de datos son superiores a técnicas de NPL con respecto a la identificación de síntomas en artículos científicos. El sistema muestra un menor desempeño en las notas médicas[32].

Problema a tratar

Enfocamos nuestros esfuerzos en la detección de signo y síntoma de una enfermedad o padecimiento, mediante la extracción de textos en publicaciones en WEB. Aplicando técnicas de procesamiento de lenguaje natural, un diccionario y expresiones regulares. Así se obtiene un conjunto de entidades de significancia médica. Se coloca una nube de palabras de representación de los síntomas menos frecuentes de la enfermedad.

METODOLOGÍA

Aplicando acopio de información de noticias y párrafos relevantes, se obtienen elementos para la construcción de un conjunto de prueba y entrenamiento.

Con el conjunto trabajado, se aplica la técnica de máquina de soporte vectorial para la clasificación de textos, donde cada párrafo se da como un entrada al proceso, apoyado de una función de base radial, Se usan agentes para la extracción de URL's y la extracción de nuevas instancias de párrafos.

Técnicas de procesamiento de lenguaje natural, como la identificación de entidades, es aplicada para la extracción de palabras y la creación de un diccionario de datos. Para ilustrar el modelo se muestra la Figura 1.

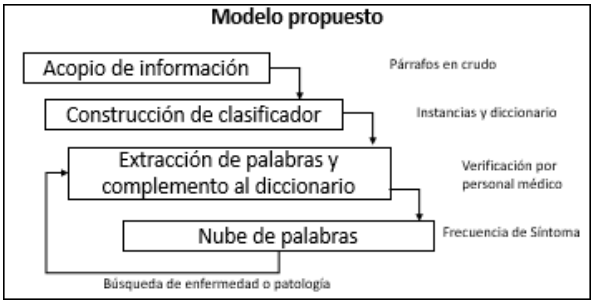

Figura 1. Modelo propuesto. En 4 etapas se describe las tareas realizadas de nuestro modelo. Cada parte del proceso produce una salida que es la entrada del otro. No se consideraron como etapas a las actividades propias de la aplicación, como la identificación de palabras con significancia médica.

Equipo utilizado

En el contexto de equipo para el proyecto se utiliza una PC para el proceso, software de código abierto y es sólo para propósitos académicos. Las características son: *Python* 3.7, con 16GB *Ram*, Ryzen 5 con 12 núcleos y un disco de estado sólido de 250GB, en una placa B450-M1, una tarjeta de

vídeo gráfica 730, fuente de poder 800 Watts. Monitor 27".

Diagrama del modelo

La Figura 1, muestra los pasos en el desarrollo. Allí se engloban las tareas en bloques, que ayudan a simplificar el entendimiento del proceso. Así entonces, se presenta como un elemento completo y entendible.

Acopio de información

Para el acopio de información se realizó el pre proceso de la Figura 2. Se usan agentes para la identificación de URL's y raspado web para párrafos.

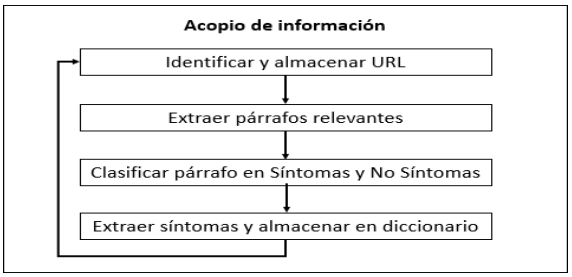

Figura 2. Adquisición de información. Aquí, se utilizan agentes para la obtención de URL's y *Web Scraping* para la extracción de párrafos de los párrafos. Se realiza una clasificación manual de textos y los términos de signo y síntoma se anexan a un diccionario de datos.

La identificación de los sitios web, podrá dar validez a la información procesada, además se ha identificado la URL, se han identificado párrafos relevantes, así mismo se ha clasificado si el texto contiene síntomas del padecimiento, ello se guarda como hoja de cálculo, después se extraen síntomas de esos textos, por lo tanto, se buscan sinónimos y se anexan a un archivo de hoja de cálculo. Así se crea un pequeño diccionario de datos. Se revisaron más de 200 artículos, desde los sitios de *Google News* y *Google.com*. Los artículos eran relacionados al COVID-19.

Construcción de un clasificador

Una vez que se pre procesó información, se procede a la construcción del conjunto de entrenamiento y prueba, para ello se etiquetaron los párrafos seleccionados en dos clases: Síntomas y No síntomas. No se revisaron aspectos como la negación o especulación. Todos los textos están en lenguaje español y no son publicaciones científicas. Se tomaron en cuenta todos los párrafos que describían algún síntoma. Un extracto de los párrafos se muestra en la Tabla 1. Usando *Sklearn* se crea un clasificador usando SVM y Ec. 1.

Adicionalmente se alimentó un diccionario de datos, con todos los síntomas detectados en estas publicaciones.

Tabla 1.- Instancias de párrafos

Núm.	Párrafo	Clase
0	A través del podcast informativo de la NDR en ...	Síntomas
1	Las quejas más comunes después de tres meses s...	Síntomas
2	Como sabes, los síntomas en la infección de co...	No Síntomas
3	Los resultados de sus investigaciones revelan ...	Síntomas
4	NaN	No Síntomas

Tabla 1. Muestra información de sólo los primeros 5 renglones, por simplicidad sólo ha mostrado parte del texto. Denota dos clases en el conjunto, Síntomas y No Síntomas, esta anotación se realiza manualmente.

Entrenamiento del clasificador

Se procedió a realizar el entrenamiento del clasificador binario. Se toman 88 instancias para el entrenamiento y 30 para la prueba. Después de procesar, se procede a la construcción de la matriz de confusión.

Extracción de palabras y complemento al diccionario de datos

Se implementa una función de tokenización, retirando las palabras de parada en *StopWords de NLK* y los símbolos de puntuación en español, se probó la lematización y la extracción de raíces. Por último se agregó la función de etiquetado, dónde se establecen los elementos sustantivos como salida más relevante del proceso, por lo tanto, se produce la entrada para la nube de palabras. Manualmente se complementa el diccionario, así se incluyen palabras no detectadas en funciones anteriores.

Nube palabras

Determinada la frecuencia de síntomas, se procede a realizar la publicación en la nube de palabras, en este caso, se toma como referencia los elementos con menos coincidencias.

RESULTADOS Y DISCUSIÓN

Resultados

El clasificador presenta una buena capacidad de discriminación, un recuerdo por arriba de la media y una exactitud dentro de lo aceptable, las métricas se muestran en la Tabla 2. La Figura 3, muestra la matriz normalizada del clasificador.

Tabla 2.- Métricas de clasificación

Ecuación	Valor
Ec. 4.- Precisión	82%
Ec. 3.- Recuerdo	80%
Ec. 2.- Exactitud	83%

Tabla 2. Muestra los resultados de la clasificación. Tomando los datos tomados de la matriz de confusión normalizada (Figura 3). Datos calculados en promedio, de “Síntomas” y “No Síntomas”.

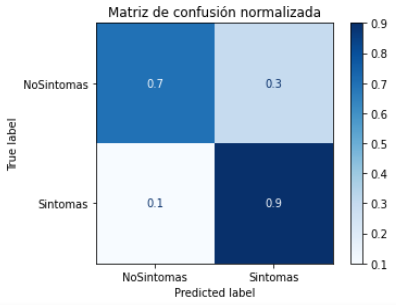

Figura 3. Matriz de confusión, los elementos de la diagonal de izquierda a derecha muestran las instancias correctamente clasificadas, es decir 90% de instancias correctamente clasificadas para la clase Síntomas y 70% para No síntomas.

Los resultados son mejores usando una base de conocimiento de síntomas, esto se muestra en la comparativa de la Figura. 4 y Figura. 5.

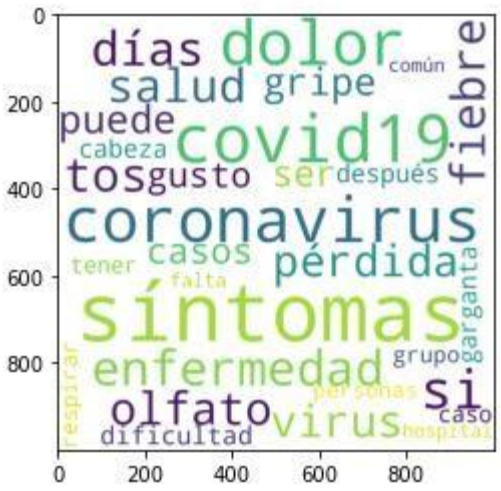

Figura 4.- Nube de palabras, utilizando etiquetas de sustantivo. Aquí se ha utilizado la capacidad de NTLK para identificar entidades de significancia médica, sin embargo algunas palabras no muestran sentido, como “salud”, “caso”, “común” y “si”.

Figura 5.- Nube de palabras, utilizando expresiones regulares. Aquí se denota a los síntomas más comunes encontrados, así logran identificarse N-gramas, sin la necesidad de la construcción manual de estos.

La Figura 4 muestra palabras que no son útiles como “común”, “ser”, “tener”, mientras que la Figura. 5. Muestra elementos más valiosos en la identificación de síntomas. La información es más útil cuando se invierte la forma de impresión a las palabras menos frecuentes Figura 6.

ISSN: 2448 - 7198

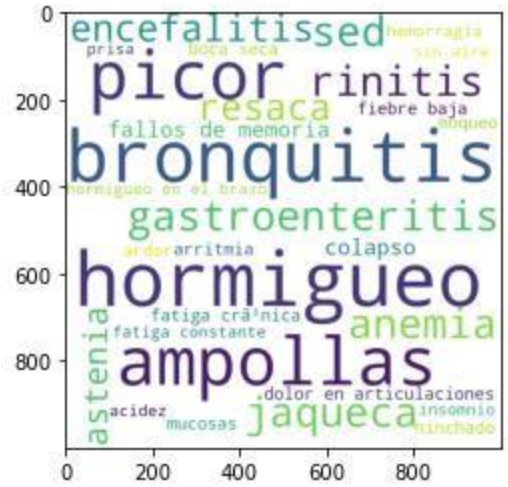

Figura 6.- Nube de palabras, utilizando expresiones regulares. Aquí se denotan las palabras menos frecuentes, pero que en las muestran síntomas graves del COVID-19 como “encefalitis”, “bronquitis”. Además, se muestran aspectos de secuelas del mismo padecimiento como “fatiga crónica”.

Discusión

Se presentó una metodología, que aborda la tarea de identificación de signos y síntomas desde una perspectiva de las publicaciones WEB, donde las noticias escritas son la entrada al proceso. El objetivo es la identificación de nuevos síntomas asociados a un padecimiento o enfermedad, que requiere que la información se presente de forma simple y contundente. Contrasta con los trabajos presentados [19, 20, 21, 22, 23, 24], debido a que nuestro trabajo incorpora un clasificador para la detección de párrafos de interés, el raspado web, el uso de un diccionario de datos y una representación visual de las entidades. Además, el uso de los agentes facilita el acceso a fuentes de información, limitando su alcance a publicaciones en español.

En comparación con [24], éste último aplica un conjunto técnicas NLP y el uso de expresiones regulares para la extracción de entidades en busca de automatizar la selección de diagnósticos médicos. Así entonces, nuestro diccionario de datos posibilita este proceso para futuros trabajos.

El modelo presenta una precisión y recuerdo satisfactorio, impactando en su capacidad de detectar párrafos que contienen información de interés, se observa que puede descubrir nuevos términos asociados a un padecimiento, así entonces se obtuvieron 704 anotaciones de síntomas. Estas anotaciones son hechas manualmente, pues requiere validar si es un término médico o una apreciación particular del autor. Así las entradas son consultadas a personal de salud.

El modelo presenta la deficiencia de no detectar N-gramas desde en etapas tempranas, derivado de no contar una lista de ellos con significado médico. Por ejemplo el Síndrome Inflamatorio Multisistémico, este diagnóstico o signo está asociado a secuelas del COVID-19. Así entonces, se aborda la tarea como una búsqueda de expresiones regulares, donde el diccionario de datos cobra relevancia.

En las Figuras 4, 5 y 6, se muestra un factor de temporalidad e incidencia. La temporalidad se da en que muchas publicaciones hablan sobre síntomas que se han convertido en conocimiento general, sin embargo en la Figura 6 se aprecian nuevos síntomas y menor incidencia en las publicaciones. Por lo tanto, el modelo sí demuestra que la tarea es posible y alcanza su propósito.

La representación de las entidades en una nube de palabras, permite visualizar de forma clara los síntomas o signos; organiza la información con base a la frecuencia, pero habilita al investigador para interpretar los datos y mostrar aquellos de interés o de circunstancias no ordinarias. Así entonces, se puede mostrar más información que la que se presenta de forma ordinaria, por ejemplo se toma la página del Gobierno de México https://coronavirus.gob.mx/, que sólo lista los síntomas de tos, fiebre y dolor de cabeza (consultada el 24 de septiembre de 2021).

CONCLUSIONES

La importancia de resaltar síntomas que aparecen en un nuevo padecimiento, es la oportunidad de que una persona puede prevenir una forma grave de la enfermedad, la posible portación de un signo o síntoma y prevenir una muerte prematura por complicaciones. Así apoya a personal médico, en la consulta o revisión.

Lo más difícil es procesar la interpretación de los usuarios, es decir, bajo su experiencia de vida ¿cómo describen los síntomas?, recordando que la sintomatología presentada es similar a diversas

enfermedades. Por la tanto, se requieren más muestras y recursos para identificar los posibles diagnósticos médicos.

Trabajos Futuros

Se visualizan muchos trabajos futuros en el ámbito académico. Será importante realizar alianzas con organizaciones de salud, que con aportaciones al diccionario de datos, mejoran la detección de nuevos síntomas. Además, el conocimiento adquirido nos lleva a plantear una ontología orientada en las relaciones de las entidades de la enfermedad.

AGRADECIMIENTOS Y/O RECONOCIMIENTOS

Se agradece al programa de estudios Maestría en Sistemas Computacionales y al PhD. Elías Ruiz Hernández del Instituto Tecnológicos Superior del Oriente del Estado de Hidalgo (ITESA) y al Instituto Tecnológico Superior del Occidente del Estado de Hidalgo (ITSOEH).

REFERENCIAS

8. Kalyan, K. S., & Sangeetha, S. (2020). SECNLP: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101, 103323.
9. Huang, Y., & Zhao, N. (2020). Generalized anxiety disorder, depressive symptoms and sleep quality during COVID-19 outbreak in China: a web-based cross-sectional survey. *Psychiatry research*, 288, 112954.
10. Kim, G. U., Kim, M. J., Ra, S. H., Lee, J., Bae, S., Jung, J., & Kim, S. H. (2020). Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. *Clinical microbiology and infection*, 26(7), 948-e1.
11. Naeem, S. B., Bhatti, R., & Khan, A. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2), 143-149.
12. Elibol, E. (2021). Otolaryngological symptoms in COVID-19. *European Archives of Oto-Rhino-Laryngology*, 278(4), 1233-1236.
13. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., ... & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10(1), 1-10.
14. Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4), 337-339.
15. Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., ... & Feng, J. (2020). Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature communications*, 11(1), 1-14.
16. Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence*.
17. Zhang, X. D. (2020). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). *Springer*, Singapore.
18. Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). *Springer*, Boston, MA.
19. Tan, X. H., Bi, W. H., Hou, X. L., & Wang, W. (2011). Reliability analysis using radial basis function networks and support vector machines. *Computers and Geotechnics*, 38(2), 178-186.
20. Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. *O'Reilly Media*.
21. Flach, P. (2012). Binary classification and related tasks. In Machine Learning: The Art and Science of Algorithms that Make Sense of Data (pp. 49-80). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511973000.004.
22. S. Russel y N. Peter. Inteligencia Artificial, un enfoque moderno, Pearson Educación, 2013, (pp. 37-38).
23. Lawson, R. (2015). Web scraping with Python. *Packt Publishing Ltd*.
24. Heydt, M. (2018). Python Web Scraping Cookbook: Over 90 proven recipes to get you scraping with Python, microservices, Docker, and AWS (pp. 218-220). *Packt Publishing Ltd.*
25. Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1), (pp. 51-89).
26. Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 26). *Springer*, Boston, MA.
27. Avila, J., & Hauck, T. (2017). Scikit-learn cookbook: over 80 recipes for machine learning in Python with scikit-learn (pp. 36-39). *Packt Publishing Ltd*.
28. Thanaki, J. (2017). Python natural language processing. *Packt Publishing Ltd.*
29. Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*.
30. Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6), 1930-1938.
31. Perkins, J. (2010). Python text processing with NLTK 2.0 cookbook. *Packt Publishing Ltd*.
32. Weisser, M. (2016). Practical corpus linguistics: An introduction to corpus-based language analysis (Vol. 43).
33. NLTK 3.6 documentation(2021, Abril 7) .Source code for nltk.tag.perceptron, reference https://www.nltk.org/_modules/nltk/tag/perceptron.html
34. López R., C. I. (2020). Marcos predicativos asociados al concepto signo y síntoma en textos sobre medicina en español. *Revista signos*, 53(103), (pp. 392-418).
35. Sánchez, M. T., & Velasco, J. A. P. (2014). También los pacientes hacen terminología: retos del proyecto VariMed. *Panace*, 15(39), 95-102.
36. Liu, R., Greenstein, J. L., Sarma, S. V., & Winslow, R. L. (2019, July). Natural language processing of clinical notes for improved early prediction of septic shock in the ICU. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6103-6108). IEEE.

37. Hussain, M., Choi, D. J., & Lee, S. (2020, January). Semantic based clinical notes mining for factual information extraction. In *2020 International Conference on Information Networking (ICOIN)* (pp. 46-48). IEEE.
38. Haas, S. S., Doucet, G. E., Garg, S., Herrera, S. N., Sarac, C., Bilgrami, Z. R., ... & Corcoran, C. M. (2020). Linking language features to clinical symptoms and multimodal imaging in individuals at clinical high risk for psychosis. *European Psychiatry*, 63(1).
39. Cruz, N. P., Maña, M. J., & Mata, J. (2010). Aprendizaje Automático versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina. *Procesamiento del lenguaje natural*, (45), (pp. 77-85).