SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Morelia

"José María Morelos y Pavón"

División de Estudios de Posgrado e Investigación
Maestría en Ingeniería Electrónica

tesis

"Sistema multimodal basado en procesamiento de lenguaje natural para la captura y consulta de información en base de datos médicas"

que para obtener el título de:

Maestría en Ciencias en Ingeniería Electrónica

presenta:
Pedro Mata Martínez
Director: Juan Carlos Olivares Rojas
Codirector: Gerardo Marx Chávez Campos
Revisor: Adriana del Carmen Téllez Anguiano
Revisor: José Antonio Gutierrez Gnecchi

Morelia, Michoacán, México – Junio 2023 – Rev 2.0

Pedro Mata Martínez: *Sistema multimodal basado en procesamiento de lenguaje natural para la captura y consulta de información en base de datos médicas,* Maestría en Ciencias en Ingeniería Electrónica, ©Junio 2023

Mesa de revisión:
Juan Carlos Olivares Rojas
Gerardo Marx Chávez Campos
Adriana del Carmen Téllez Anguiano
José Antonio Gutierrez Gnecchi

Localidad:
Morelia, Michoacán, México

Impresa:
Junio 2023

Abstract

The language we use is a form of communication through the use of words, this allows us to understand and learn more about the world around us. natural languages are those that people use to communicate with each other, regardless of language, plus it is possible to transmit the same message using different types of statements or words, whether they are vague or precise. Natural language processing encompasses what a computer needs to understand the natural language that people use, and thus generate responses according to the needs with which they are used. This paper documents a problem about the data capture in the area of health, more precisely speaking, diabetes; because when the questioning and physical examination is performed on a patient, the doctor cannot perform tasks simultaneously, impeding better patient treat. hat being said, the main objective of this research is the capture and retrieval of information for filling in fields of medical records of patients with diabetes using analysis of language processing natural to recordings of medical consultations.

Resumen

El lenguaje que usamos es una forma de comunicación mediante el uso de palabras, esto nos permite comprender y aprender más acerca del mundo que nos rodea. Los lenguajes naturales son aquellos que la gente usa para comunicarse entre sí, independientemente del idioma, además de que es posible transmitir un mismo mensaje utilizando diferentes tipos de enunciados o palabras, ya sean estas vagas o precisas. El procesamiento de lenguaje natural engloba lo que una computadora necesita para comprender el lenguaje natural que usan las personas, y así generar respuestas acordes a las necesidades con las que sean utilizadas. En este escrito se documenta una problemática acerca de la captura de datos en el área de la salud, más precisamente hablando, de la diabetes; debido a que cuando se realiza el interrogatorio y la exploración física a un paciente, el médico no puede realizar tareas de manera simultánea, impidiento dar una mejor atención al paciente. Dicho esto, el principal objetivo de esta investigación es la captura y recuperación de información para el llenado de campos de historiales médicos de pacientes con diabetes mediante el análisis de procesamiento de lenguaje natural a grabaciones de consultas médicas.

Glosario

Índice general

Capítulo 1

Protocolo

1.1 Introducción

En el año 2022 mediante la Encuesta Nacional de Salud y Nutrición (Ensanut) en México se registraron 12.4 millones de personas con diabetes, casi un 10 % de la población total del país, por lo cual la implementación de módulos para la atención de esta enfermendad ha crecido en la última década, tanto así como la demanda de su atención y de medicamentos para su tratamiento.

Debido al crecimiento de los sistemas de información se está teniendo problema con la búsqueda de la información. Particularmente es complicado la captura de información ya que se hace de forma manual por el médico que tiene que llenar los campos y por lo tanto, deja de prestar atención al paciente a quien se le está realizando la consulta.

El procesamiento de lenguaje natural (NLP por sus siglas en inglés), es un conjunto de métodos y técnicas de la computación para el análisis y la interpretación del lenguaje humano. Que mediante el uso de Inteligencia Artificial (A.I.), el programa de análisis sea capaz de otorgar una respuesta esperada a cierto tipo de preguntas dentro de un área como si fuera una persona con quien conversara un usuario.

El NLP puede ser usado actualmente para una gran cantidad de tareas gracias al desarrollo de la tecnología y en especial de las A.I., tales como la extracción de información, traducciones entre distintos idiomas, recuperación de información, la minería de datos, sistemas de búsqueda de respuestas (que será en lo que se basará esta investigación), entre otras.

Debido a que existen tantos campos en los cuales se puede especializar un programa de NLP, este plantea una dificultad para el desarrollador, ya que, cada sistema a analizar utiliza distintos tipos de métodos y técnicas para la búsqueda, análisis, interpretación y respuesta para el usuario. Por ejemplo, para la búsqueda de imágenes se utiliza una I.A. entrenada previamente para los resultados encontrados que coincidan con la solicitud se desplieguen tal y como el usuario haya pedido; para la búsqueda de datos estructurados como es en bases de datos, se requiere de una sintaxis especial conocida como

lenguaje de programación, que pueda interpretar tanto el programa de NLP como el programa que administra la base de datos.

En el caso de esta investigación, será centrado en el uso del NLP para la captura de audio de una cosulta médica, para poder ser analizada porteriormente, y mediante la interpretación de los diálogos, clasificar la información de la interrogación y exploración médica que se realiza sobre el paciente, almacenarla en una base de datos para obtener futuros diagnósticos que apoyen a la interpretación de un correcto análisis de la consulta. Lo que así permite el desarrollo más especializado de la investigación, para determinar que lenguaje de programación utilizar, que tipo de interfaz para el usuario desarrollar, qué métodos y técnicas implementar, si es posible utilizar bases preliminares de terceros, que tipo de salidas serán desplegadas para el usuario, entre otros.

1.2 Estado del arte

La historia del NLP se puede remotar desde la década de los 40s. Cuando las Máquinas de Traducción, Machine translations (MT), los cuales fueron los primeros dispositivos en los que se realizaron aplicaciones relacionadas al lenguaje natural.

Después que se empezó a acuñar el término de la inteligencia artificial en la década de los 50s, donde la invención de la Máquina de Turing, creada por Alan Turing, la cual realiza una prueba para verificar si una máquina se podría considerar como una máquina inteligente, en otras palabras, si mediante las entradas que se le aporten a la máquina, el comportamiento o resultados de esta sea similiar o hasta indistinguible del ser humano. [1]

Al momento de definir lo que es un lenguaje natural, en [2] se dice que un lenguaje natural es aquel que ha evolucionado con el tiempo para fines de la comunicación humana. El uso de las palabras para de manera oral o escrita con el fin de comunicar un mensaje. Los lenguajes están definidos por normas, y por lo tanto, se rigen estrictamente bajo estas.

Las aplicaciones que se le puede dar al NLP son muy variadas, algunas de las aplicaciones para las cuales se usa son: tutores inteligentes, traducción automática de idiomas, obtención de resúmenes, resolución cooperativa de problemas, recuperación de información, extracción de información y reconocimiento de voz [2]. Siendo estas últimas dos aplicaciones los tópicos en los cuáles se basará el proyecto para realizar la distinción de voces entre el médico y el paciente, y la extracción de la información relevante para la documentanción del historial médico del paciente.

Para realizar el reconocimiento de voz, en [3] usa un Modelo Oculto de Markov (MOM), el cual usa un modelo generativo, en otras palabras, una secuencia de razonamiento extendido de Modelos Bayeasianos (MB), en el cual, mediante un conjunto de evidencias dependientes de una secuencia de estados ocultos se pueda predecir las salidas de próximos estados mediante un entrenamiento de los anteriores. Para aplicarlo al reconocimiento de voz, en el análisis de las formas de onda de las palabras pronunciadas (evidencias), deben de encajar con la secuencia de los fonemas individuales (estados ocultos) que más se parezcan al producirlo.

En [4] se realiza un análisis sobre las noticias basado en la pureza textual y el análisis de la lingüística, así como explorar en la estrcutura de los artículos de las noticias y la influencia de la redacción mediante la Pirámide Invertida de las 5W1H (Preguntas con W y H en el idioma Inglés). Qué (What), quién (who), cuándo (when), dónde (where), por qué (why) y cómo (how). Con el objetivo de revisar la qué sucesos ocurren, quién los está realizando, cuándo sucedieron, en qué lugar, por qué y cómo; para distinguir las secciones del texto escrito y cómo se relacionan unas con otras.

Usando un sistema evaluación de modelos entrenados para la recuperación de acentos y mayúsculas en enunciados en Catalán y Gallego [5], mediante la tokenización de las palabras dadas como entrada al sistema entrenado. Como se mencionó anteriormente, en este sistema se da un enunciado de entrada en minúsculas y sin los acentos que debería tener según las normas del idioma, se separa en tokens que son mandados a un sistema preentrenado, se condicionan las salidas y se obtiene una capa de clasificación de tokens con los tipos de correcciones que se deben implementar como el cambio a mayúsculas o minúsculas, adición de comas (,), puntos (.), dos puntos(:), signo de interrogación (?), signo de admiración (!).

Para un desarrollo de un chatbot o asistentes conversacionales, los cuales son propgramas informáticos capaces de mantener un conversación hablada o escrita, como si se tuviera dicha conversación con una persona. En [6] se pudo entrenar un chatbot con un conjunto de conversaciones separadas en tokens de los diálogos de 4 distintas series televesivas y cinematográficas con un DialoGPT basado en el NLP. Lo cual nos permite tener una nueva manera de interpretación del texto por medio del lenguaje natural.

Con el fin de facilitar la compresión de términos y temas relacionados a la salud por parte de la sociedad, mediante la creación de un recurso parelelo sintético en español para el entrenemiento y validación de métodos de simplificación de cuestiones sanitarias. Se muestra como en [7] se da una oración o párrafo con terminología médico relacionada a la salud, con tal de obtener como resultado una oración o un conjunto de estas que los simplifiquen para que las demás personas que no estén relacionadas con esos términos las entiendan con mayor facilidad. Por lo cual, al ser el paciente quien va a otorgar la información, debería ser posible analizar los diálogos para su recuperación e interpretación a términos sanitarios.

En [8] utiliza un método de análisis del NLP para la detección de errores en notas médicas en el idioma español. Primeramente se encuentra el error, es decir, si alguna palabra está mal escrita, para después generar un conjunto de palabras sugeridas con las que se debe reemplezar. Esto debe de tomar en cuenta a que hay varias maneras en las que una palabra sea mal escrita, y que hay otras palabras que pueden asemejarse a esta sin ser la correcta. Esto puede ayudar a la investigación de manera que al momento de realizar la captura del audio y la transcripción a texto para ser analizado, dicho texto pueda hacerlo de manera incorrecta por uno de los siguientes factores: falta una o más letras, tiene letras demás, o que la palabra pronunciada ni siquiera exista en el lenguaje español.

Mediante el uso de redes neuronales recurrentes y redes de memoria, se da a lugar a las redes neuronales de memoria, las cuales poseen una memoria estructurada para el almacenamiento y para la codificación a corto y largo plazo, con el fin de una vez dados los textos a analizar, se vectorizan, se someten a un entrenamiento y validadición para comparar con una respuesta predicha y conocer los índices de error, como se menciona en [9]. Esta investigación mencionada se centra en la clasificación de textos basados en el Machine Learning (M.L.), con la capacidad de analizar los textos que se le introducen al sistema en modo de pregunta y obtener como respuesta una oración tokenizada en orden de tal manera que su gramática sea correcta en el idioma español.

La manera más óptima para el NLP es mediante el método de la vectorización, dicho método es implementado en [8], [9], [10], [11], [12] y [13]. Esto es la generación de vectores one-hot encoding de los vocabularios a implementar, es decir, por cada palabra existente en el sistema se creará un vector de tantos renglones como palabras a usar, pero cada palabra tendrá un único valor de 1 en su respectivo vector, a consecuencia los demás índices del vector serán marcados con 0. Debido a que los sistemas de redes neuronales no pueden analizar texto directamente.

Existe una gran relevancia del uso de las redes neuronales aplicados a NLP, a causa de que la cantidad de vectores a analizar sean tan grandes que afecte de manera negativa a su eficiencia y eficacia, por lo tanto, es necesario crear un sistema que pueda reducir la cantidad de vectores con

valores que representen las similitudes entre los distintos vectores originales como en los trabajos de [9], [8], [9], [14] y [15].

Para esta investigación la rama de estudio como ya se ha mencionado es la medicina, en específico la diabetes, por lo tanto es importante tener en cuenta la terminología médica aplicada a notas médicas y consultas a pacientes diabéticos. En estos casos para el NLP es necesario contar con un conjunto de asimilaciones para las posibles palabras a interpretar por el médico o por el paciente en cuestión, así como en [7], [16], [17], [18] y [19]. Sin restarle importancia al tipo de almacenamiento de datos de historiales médicos en bases de datos como en [20] y [21], ya que la finalidad de este proyecto es la recuperación, clasificación y alamacenamiento de campos de historiales médicos de pacientes con diabetes.

1.3 Solución propuesta

Método propuesto

Para el problema descrito se optará como propuesta de solución la creación de un programa que pueda guardar la conversación de una consulta médica enfocada en el tópico de la diabetes, para posteriormente analizarlo, obtener la distinción de las voces del paciente y el médico en cuestión, para mediante el código de NLP desarrollado, recuperar la información relevante de la consulta con tal de obtener un historial médico.

Hipótesis

El desarrollo de un sistema multimodal basado en procesamiento de lenguaje natural permitirá una captura y búsqueda más rápida de los datos en base de datos médicas.

Objetivos

General

Investigar métodos de procesamiento de lenguaje natural para la implementación de un sistema multimodal que permita agilizar la captura y búsqueda de información en base de datos médicas

Específicos

- Obtener el listado de la información relevante para una consulta médica enfocada en diabetes por un médico especialista en el área.
- Separar en secciones los tipos de información para un historial médico de un paciente generado por una consulta médica enfocada en la diabetes.
- Crear una base de datos en MySQL que pueda almacenar todos los campos necesarios relacionados a un historial médico de un paciente con diabetes.

- Crear un código en Python que pueda realizar la distinción de voces entre dos distintas personas con tal de diferenciar entre los diálogos del paciente y del médico especialista en diabetes.
- Crear un código en Python que pueda almacenar una conversación de audio para posteriormente transcribirla a un archivo de texto.
- Crear un código en Python con la capacidad de analizar el texto obtenido con NLP con la finalidad de recuperar la información relevante para el historial médico.
- Almacenar la información relevante para el historial médico en la base de datos.

1.4 Materiales y métodos

Metodología general

El proceso que se seguirá de manera general para el proyecto de investigación se muestra en la Figura 1.1. Para conseguir el listado de la información relevante para la generación de un histiorial médico con enfoque a la diabetes se optará por obtener a una cita con un médico especialista en el área que pueda brindar al estudio de la investigación todos aquellos aspectos que toma en una consulta médica y cómo estos son analizados. Para este caso será el módulo del Centro de Atención a la Diabetes del Instituto Mexicano del Servicio Social (CADIMSS).

Después de analizar la estructura de las bases de datos hospitalarias, se procederá a crear un modelado de la conversación entre el médico y el paciente, en la cual se estudiarán los diálogos al momento de realizar la consulta, con tal de conocer las preguntas y respuestas que son de utilidad para llenar los campos de un historial médico de un paciente con diabetes.

Con el modelo ya definido, el siguiente paso es estudiar directamente las conversaciones de las consultas médicas para generar un vocabulario que tendrá las posibles palabras que se pueden llegar a comentar en una consulta con un médico especialista en diabetes. Dicho vocabulario deberá de contar con artículos determinados, indeterminados, preposiciones, entre otras palabras básicas para la gramática del idioma español, pero deberá contar principalmente con terminología médica y frases o palabras similares que se puedan asimilar a estos términos.

Por parte del hardware (HW), se usará un dispositivo embebido que pueda soportar la codificación de un lenguaje de programación para el NLP, entrenamiento de redes neuronales, manipulación de base de datos y captura de audio. El dispositivo que cumple con esas características es el Raspberry Pi4, el cuál es un dispositivo embebido basado en el sistema operativo de GNU/Linux. Para la captura de audio se propone usar un micrófono que tenga la suficiente sensibilidad para escuchar las voces que se encuentren dentro del cuarto donde se estén realizando las consultas, y la claridad para distinguir las palabras que se pronuncien durante estas.

Para el caso del software (SW), se dividirá en 3 partes: la creación de la base de datos, la creación de una página web que despliegue de manera visual y ordenada la información del historial médico, el código del algoritmo que permita hacer las tareas de captura de audio, transcripción a texto, tokenización, vectorización, compactación, apliación de redes neuronales y manipulación de bases de datos.

Después con los modelos y vocabularios obtenidos se realizará un entrenamiento en redes neuronales para que el algoritmo regrese los resultados que se esperan, para este caso la recuperación de información para el llenado de campos de un historial médico. Para posteriormente ser puestos a prueba

con consultas realizadas por doctores a pacientes con diabetes y contar la cantidad de porcentaje de aciertos que generó al momento de llenado de los campos del historial, para verificar si hay información perdida o se llene algún campo con una información incorrecta.

La creación de la base de datos para almacenar todos los campos de un historial clínico se realizará en MySQL, debido a que es una aplicación de Open Source Software (OSS), de código abierto, es decir que su código para ser públicamente accesible y su manipulación no representa una alta dificultad.

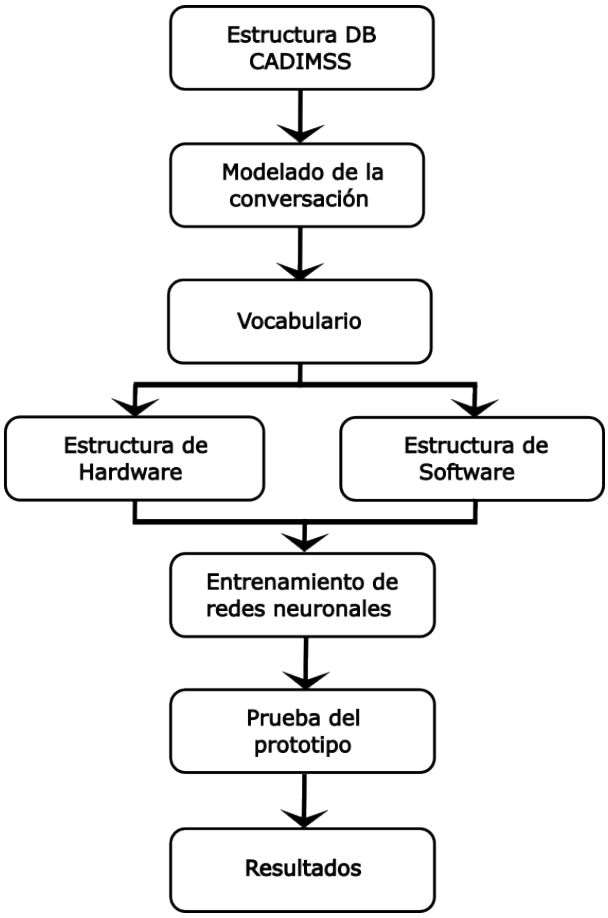


Figura 1.1: Metodología para el desarrollo de la implementación de NLP en la captura y recuperación de información de consultas hospitalarias.

Para el proceso de la aplicación del algoritmo basado en NLP se pretende seguir la siguiente estructura Figura 1.2, la cual cuenta con los siguientes pasos:

1. Captura del audio de la consulta por medio de un micrófono que tenga la suficiente sensibilidad para escuchar tanto la voz del médico como del paciente en la habitación donde está siendo atendido.
2. Una vez guardado el audio de la consulta, seccionar las partes donde se encuentre voz para transcribirlo a texto y poder ser analizado por el código de programación.
3. Cada palabra será seccionada por separado en elementos conocidos como tokens, cuyo proceso se llama tokenización, y a cada palabra distinta se le dará un valor numérico para ser analizado

por una red neuronal.

4. Cada token será transformado en un vector, que tendrá tantos índices (elementos) como palabras en el vocabulario, donde cada uno de ellos tendrá un valor de cero, a excepción de uno, en el cual en cada palabra ese valor de uno será en un índice diferente.

5. Los vectores serán sometidos a un análisis de A.I. conocido como redes neuronales. De entrada a las redes se darán los vectores usados del vocabulario encontrados en las grabaciones, ya que estos serían demasiados y demasiado grandes para que el programa trabaje de la manera más óptima. Dicho esto, serán reducidos a un conjunto de vectores más pequeños que puedan guardar la similitud entre las palabras utilizadas en la consulta con el vocabulario; este proceso se conoce como compactación.

6. Una vez obtenidos los vectores con los que se puede trabajar de manera eficiente, serán analizados por el logaritmo del programa creado, para recuperar y ordenar la información necesaria para el llenado de campos del historial médico.

7. La información recuperada será almacenada en una base de datos que contenga los campos de historial médico.

8. Se desplegará la información de manera visual en una página web para su mejor interpretación, además de modificar manualmente los datos que se encuentren en la base de datos.

   Es importante tomar en cuenta lo siguiente: Ya existen herramientas de ayuda conocidas como Interfaz de Programación de Aplicaciones (APIs) que realizan parte del trabajo de investigación, con el fin de darle un mayor enfoque e importancia a la prorgamación del código de NLP. Dichas tareas que pueden ser facilitadas por las APIs son:

   - Captura y grabación de audio
   - Transcripción de voz texto
   - Tokenización
   - Vectoriazación

1.5 Resultados preliminares

Como se ha mencionado anteriormente, el enfoque de la investigación se va a dar principalmente al NLP, debido a la existencia de APIs que apoyan el trabajo del proyecto. Para conocer que tipo de resultados se esperan del proyecto es necesario saber con tipo de información se está trabajando. Por lo cual, de primera instancia se realizó una entrevista a médicos de las localidades de Zamora, Michoacán, México y Morelia, Michoacán, México. Lo que se busca con la entrevista es conocer los tópicos y los porcentajes de estos que son hablados en una consulta médica. En la siguiente tabla Tabla 1.1 se muestra un promedio de los porcentajes de los tópicos por localidad, con un total de 23 médicos entrevistados en Morelia y 16 en Zamora.

La tabla Tabla 1.1 nos indica que la mayor parte de las conversaciones de las consultas médicas son dedicadas al análisis del paciente, debido a esto, se puede crear un clasificador de los tópicos de las consultas grabadas para el proyecto de investigación, y así, verificar que las palabras de cada tópico estén en donde corresponden y hacer caso omiso de aquellas que no formen parte de un historial médico.

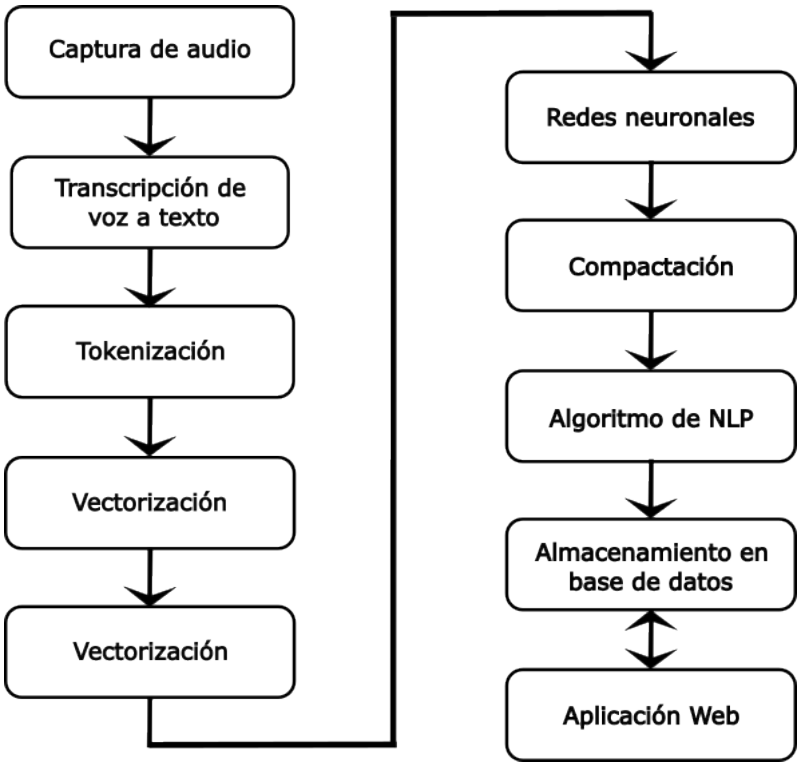


Figura 1.2: Implementación del algoritmo basado en NLP.

Ciudad	Doctores	Sintomatología	Antecedentes	Exploración	Contexto	Total
Morelia	23	23 %	37 %	26 %	14 %	100 %
Zamora	16	32 %	28 %	25 %	15 %	100 %

Tabla 1.1: Porcetaje de tópicos hablados en consultas por médicos de las localidades de Zamora y Morelia.

Para el uso de las APIs se ha probado con lo que es la captura de audio, transcripción a texto y tokenización. Para el caso de la tokenización lo que se requiere es contar con un texto, convertido a lo que para un lenguaje de programación pudiera interpretar: una cadena de caracteres. Para el caso de prueba se usó como texto una descripción de una nota médica de un paciente que recién ha sido diagnosticado con diabetes en Python, pero además de procesar la separación de dichas palabras, se buscaron palabras clave relacionadas a un historial médico de un paciente con diabetes. Se muestra el resultado de la prueba en Figura 1.3.

Para el caso de la captura de audio se ha usado una librería en Python que permite en el audio escuchado por una o más personas dentro del rango de sensibilidad de un micrófono sea salvado en un archivo de comprensión digital para audio y video, mayormente conocido como extensión .mp3 y puede ser usado para posteriormente transcribir las voces grabadas en texto que pueda ser analizado por el código de Python.

```
Texto Completo:

 paciente de 60 años que acude la farmacia por sentirse muy cansada y con una pérdida de peso importante. Últimamente
ntervenciones: tras medirle en consulta farmacéutica la glucemia y obtener como resultado 531 mg/dl, procedemos a real
 consulta y la derivamos a urgencias.

ImpresiÃ³n de todos los tokens del texto:

a', 'urgencias']

 Tokens Totales:  86

Búsqueda de palabras: hemoglobina, glicosilada y glucemia

Lista Final eliminando las palabras vac ias (no relevantes):

 ['paciente', '60', 'años', 'acude', 'farmacia', 'sentirse', 'cansada', 'pérdida', 'peso', 'importante', 'Últimamente', 'r
efiere', 'boca', 'seca', 'va', 'baño', 'La', 'paciente', 'usa', 'tratamiento', 'crónico', 'levotiroxina', '100', 'mcg', 't
ratar', 'hipotiroidismo', 'Intervenciones', 'tras', 'medirle', 'consulta', 'farmacéutica', 'glucemia', 'obtener', 'resulta
do', '531', 'mgdl', 'procedemos', 'realizarle', 'hemoglobina', 'glicosilada', 'obteniendo', 'resultado', '109', 'Se', 'rea
liza', 'informe', 'escrito', 'reflejan', 'resultados', 'obtenidos', 'consulta', 'derivamos', 'urgencias']

Total de Tokens sin Stopwords:  53

Se encontró la palabra 1
Se encontró la palabra 2
Se encontró la palabra 3
```

Figura 1.3: Resultado de la tokenización de nota médica de un paciente diagnosticado con diabetes con búsqueda de palabras clave.

1.6 Alcance y limitaciones

Unas de las limitantes para el proyecto tiene que ver las características de la voz humana son la intensidad de la voz, La amplitud de una señal de audio de una persona depende de la intensidad con la que se genere la voz, esto por ambos lados puede afectar de manera negativa al momento de realizar el análisis debido a la incapacidad de detección de palabras; y el timbre, que es lo que determina la voz de una persona, lo que permite a otros poder distinguirla de los demás, de igual manera que la entonación, si la voz del paciente y del médico tienen timbres parecidos, podría generar problemas al momento de distinguir las voces de dichos entes.

La identificación de los lenguajes de programación que tengan alguna librería especializada en los siguientes temas: grabación de audio, transquipción de audio a texto, NLP, bases de datos (MySQL); y de las librerías más optimas para el desarrollo de la investigación, ya que mediante una búsqueda realizada se ha encontrado un significativa cantidad de librerías que pueden realizar las tareas necesarias para la investigación.

El proyecto de investigación se pretende enfocar en los pacientes que acuden al médico especialista por el padecimiento de la diabetes, no se aspira a llegar a realizar un historial médico de una consulta de medicina general o a cualquier otro tipo de especilidad dentro del área, en virtud de que los campos a analizar aumentarían de tal manera que afectaría a la eficiencia y eficacia del NLP.

Los criterios a tomar en cuenta en la investigación del proyecto son: la creación y manipulación de la base de datos, la creación de un programa que pueda guardar el audio de una consulta médica, distinguir entre las voces de dicha conversación, transcribir el audio a texto, analizar los diálogos y extraer la información relevante para ser almacenada en los distintos campos de un historial médico.

1.7 Cronograma de actividades

Para la organización de las actividades para llevar a cabo el desarrollo del proyecto de la investigación serán segmentadas en quincenas, es decir, periodos de 15 días, tomando como punto de partida el mes Junio del año 2023, y finalizaría 28 semanas después, el cual corresponde con el final del mes de Julio del año 2024, tomando en cuenta la entrega de la documentación necesaria. En la tabla **??** se puede apreciar qué actividades serán a las que se les dedicará el tiempo en cada periodo que ha sido anteriormente descrito.

Actividades	Periodo de Junio a Diciembre de 2023													
	Jun		Jul		Ago		Sep		Oct		Nov		Dic	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Redacción de tesis														
Implementación de hardware														
Investigación de APIs para captura de voz														
Investigación de métodos para transcripción a texto														
Implementación de métodos de tokenización														
Investigación de métodos para vectorización de tokens														
Investigación de Inteligencia Artificial														
Investigación de Redes Neuronales														
Investigación sobre métodos de implementación de NLP														
Creación del algoritmo NLP														

Actividades	Periodo de Enero a Julio de 2023													
	Ene		Feb		Mar		Abr		May		Jun		Jul	
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Redacción de tesis														
Creación del algoritmo NLP														
Creación y actualizaciones del almacenamiento en bases de datos														
Creación y actualizaciones de aplicación web														
Prueba de programa NLP en consultas médicas para recuperación de datos														
Redacción de artículo científico														
Modificaciones finales														
Entrega final de documentación														

Tabla 1.2: Cronograma de actividades para el desarrollo del proyecto de Sistema multimodal basado en procesamiento de lenguaje natural para la captura y consulta de información en base de datos médicas en el periodo de junio 2023 - julio 2024.

1.8 Conclusiones

Con el fin de expresar las conclusiones, se compondrá por cuatro partes. La primera es el importante enfoque que se le debe dar el desarrollo del NLP. La segunda es la importancia de la elección del lenguaje de programación en el que se pretende desarrollar el código de NLP. La tercera es el desarrollo de las aplicaciones y las bases de datos para ser manipulados por el algoritmo creado. La cuarta es la verificación del mejor sistema embebido para el desarrollo de HW.

La investigación determinó que el problema debe ser centrado en la programación del NLP, debido a que ya existen herramientas de apoyo como APIs que facilitan aspectos como la transcripción del audio a texto, tokenización, vectorización y el reconocimiento de voz, por lo cual, no se les pretende dedicar más tiempo de estudio que al NLP.

El NLP es una rama de la A.I., enfocada al Machine Learning (M.L.), pero se encuentra dividida en diferentes tipos de análisis de procesamientos de lenguaje, por lo que se debe determinar cuál es el más adecuado para llevar a cabo el proyecto de investigación.

Python ha demostrado ser el software más adecuado para poder realizar el algoritmo de NLP, a comparación de otros lenguajes de programación como Java, debido a la cantidad de bibliotecas y herramientas diseñadas para su desarrollo. Además, de que existen modelos de redes neuronales para la compactación de vectores que pueden ser utilizados como un punto de inicio de la programación en el lenguaje de Python, en lugar de desarrollar uno desde cero.

Las bases de datos al ser sistemas de almacenamiento de información deben ser estructurados de tal manera que los campos estén bien definidos sin que intervengan unos con otros, esto es de vital importancia ya que los datos almacenados en las bases de datos médicas no deben de mezclar ningún tipo de información con algún otro, almacenar en campos equivocados o que la información no sea colocada en el paciente que está siendo atendido.

Existen muchos sistemas embebidos que pueden correr códigos de diferentes lenguajes de programación, sin embargo, para el caso del proyecto de investigación es necesario que se pueda correr e interpretar un lenguaje de programación que pueda cumplir con todos los aspectos anteriormente marcados, de lo contrario, esto presentaría una pérdida de recursos y tiempo para el desarrollo de la investigación.

Capítulo 2

Marco Teórico

2.1 Procesamiento de Lenguaje Natural

2.1.1 Definición

Se puede entender al NLP como la unión del procesamiento computacional de la A.I. con la lingüística computacional, con el fin de que las computadoras tengan la capacidad de comprender textos, palabras y sílabas así como lo puede hacer una persona. Por lo cual se puede entender al NLP como una subrama de la A.I. en la que se pueden aplicar modelos estadísticos, de Machine Learning (M.L.) y Deep Learning (D.L.).

Para que una computadora pueda interpretar un lenguaje escrito o hablado es necesario que cuente con alguna aplicación del campo de la informática, para lo cual se ha desarrollado la lingüística computacional, con el fin de analizar, sintetizar y comprender el lenguaje.

2.1.2 Inteligencia Artificial

Como se ha mencionado anteriormente, para que una máquina o dispositivo se considere como una máquina inteligente es necesario que esta pueda otorgar resultados de manera similar o igual a como lo podría hacer una persona; por lo cual se ha estado desarrollando la aplicación de métodos y técnicas para que las computadoras sean capaces de realizar tareas así como lo es capaz la inteligencia humana.

El avance de la A.I. en la última década ha sido tan considerable que ahora hay trabajos que anteriormente eran realizados por personas expertas en algún campo o tema, han sido reemplazados por programas de A.I. gracias a su capacidad de análisis, velocidad de procesamiento y reducción de costos.

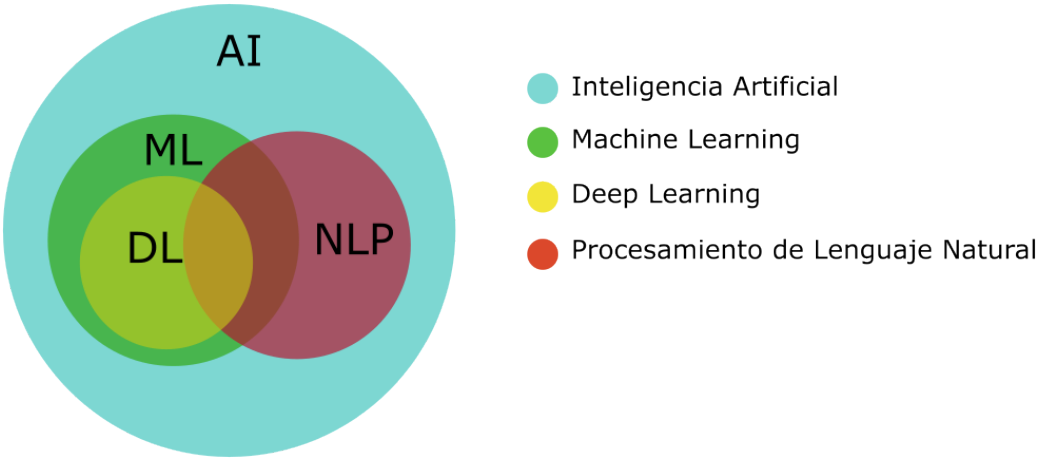


Figura 2.1: Relación del NLP con la A.I. y sus subconjuntos o subramas.

2.1.3 Machine Learning

La traducción literal del M.L. sería aprendizaje máquina, lo cual no está muy alejado de su definición, debido a que cuando se desea aplicar el M.L. a una computadora, ésta debe de comportarse de manera similar a como lo haría una persona en cuanto a aprendizaje se refiere. Para que se haya considerado que algo se ha aprendido es necesario ser capaz de predecir resultados según las características dadas, por lo tanto el M.L. es un conjunto de algoritmos y reglas que le permiten a una computadora identificar y aprender patrones de datos con el fin de realizar predicciones.

Métodos de aprendizaje del Machine Learning

Existen diferentes métodos para que una A.I. pueda predecir resultados de los patrones que se le hayan dado anteriormente: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado.

1. Aprendizaje no supervisado

Para poder enseñar a una máquina a realizar predicciones es necesario que sepa como distinguir entre los diferentes grupos que se están analizando mediante la o las respuestas esperadas, por decir un ejemplo, si pertenece al grupo A o al grupo B. Dados un conjunto de datos puestos de manera gráfica se pueden apreciar las posiciones de manera cartesiana en una, dos o tres dimensiones, donde a simple vista de una persona es posible diferenciar cuáles resultados pertenecen a un grupo y cuáles al otro respectivamente, pero a una máquina hay que diseñar un método de separación y agrupamiento, y a esto se le conoce como clustering.

Los métodos de clustering normalmente es seccionado o separado mediante una función matemática, como lo pueden ser una línea recta, una curva o un cuerpo cerrado (un ejemplo podría ser una elipse), en la que dicha funsión delimita si los puntos de cada una de los eventos analizados pertenecen a un grupio u a otro.

Otro método para el aprendizaje de una máquina es por medio de la asociación. Dadas un conjunto

de datos a analizar pueden ser muy similares entre sí, pero deben ser separados por las características con las que cuentan, a diferencia del agrupamiento que se basa en buscar el valor numérico de las variables a medir, la asociación busca las diferencias del conjunto de datos.

1. Aprendizaje supervisado

Este método para la clasficiación de resultados, es decir, la predicción de salidas requiere de una etiquetación de los datos a analizar. Dada una serie de datos previamente clasificados y etiquetados por un experto en el tema, son puestas a ser analizados por la A.I. mediante una serie de pruebas conocida como entrenamiento.

2.1.4 Deep Learning

El D.L. es una subrama a su vez del M.L. que está enfocada al aprendizaje de la máquina lo más acertado a como lo haría una persona, pero mediante el uso de las redes neuronales. Una red neuronal es una imitación de las neuronas del cerebro humano la cual se encarga de manejar datos de manera numérica, pero también existen redes neuronales para analizar imágenes que se explicarán más adelante.

2.2 Diabetes

2.2.1 Definición

Es una enfermedad en la que los niveles de glucosa en la sangre, más comunmente mencionado como azúcar, están por encima de lo normal. La mayor parte de los alimentos que las personas consumen se transforman en glucosa para ayudar al cuerpo a tener energía, pero cuando el cuerpo no funciona correctamente para la absorción de la glucosa es debido a que el cuerpo de esa persona sufre de diabetes.

El órgano encargado de para de la glucosa que se encuentra en la sangre es el páncreas, el cual crea una enzima conocida como insulina que le permite el transporte y absorción de la glucosa a las células del cuerpo. Por lo tanto si un cuerpo tiene diabetes es por el acumulamiento de la glucosa en la sangre por la falta de producción de insulina o una generación de una resistencia a la insulina.

Figura 2.2: Imágen respresentativa del órgano páncreas, su localización y tamaño aparente.

2.2.2 Tipos

A lo largo de la historia desde el descubrimiento de esta enfermedad se han detectado tres tipos de diabetes: diabetes tipo 1, diabetes tipo 2 y diabetes gestacional:

- Diabetes tipo 1

El cuerpo no es capaz de producir insulina.

- Diabetes tipo 2

El cuerpo genera una resistencia a la insulina producida por el cuerpo o la insulina producida no es suficiente para que pueda ser aprovechada por las células del cuerpo. Cabe destacar que es el tipo de diabetes más común entre las personas que la padecen.

- Diabetes gestacional

Es un transtorno que se produce durante el embarazo afecta la forma en las células del cuerpo absorban la glucosa de la sangre. Controlar la glucosa de la sangre es importante debido a que puede generar problemas durante el parto, así como generar un aborto o que el infante nazca con diabetes. Generalmente después del parto los niveles de glucosa en la sangre de la madre regresan a la normalidad, sin embargo, aumenta la posibilidad de que se desarrolle una diabetes de tipo 2.

2.2.3 Parámetros

Para poder conocer la salud del cuerpo se realizan estudios médicos de laboratorio, y en para el caso de los niveles de glucosa en sangre es necesario el estudio conocido como química sanguínea. La química sanguínea muestra en sus resultados la urea, creatinina, ácido úrico, colesterol, triglicéridos, colesterol HDL, bilirrubina, y el necesario para el conocimiento del estado de diabetes de los pacientes, la glucosa. La glucosa nos permite saber qué niveles de azúcar en sangre tiene la persona que se ha realizado los análisis de laboratorio, pero para determinar si un paciente es diabético con un menor índice de error, es mediante la hemoglobina glicosilada.

La hemoglobina glicosilada es un estudio de sangre que nos indica el nivel o volumen promedio de glucosa en sangre durante los últimos meses, es decir, que es necesario el transcurso de un tiempo determinado desde que se conozca que tiene niveles de glucosa superiores a los normales. Realizar esta prueba es muy útil para determinar si un paciente es prediabético, si sufre de diabetes tipo 1 o diabetes tipo 2, y realizar y mejor control de seguimiento.

Para que estos salgan con valores esperados a su estado normal, es necesario que los estudios se hayan realizado mientras el paciente se haya encontrado en ayuno, ya que ingerir alimentos o bebidas con carbohidratos podrían alterar dichos niveles.

70-100

5.7 6.5

Es posible que una persona

2.2.4 Sintomatología

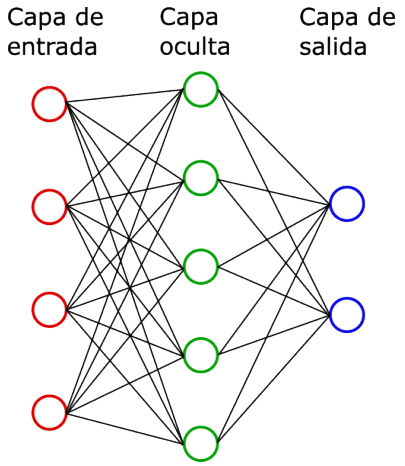


Figura 2.3: Diagrama representativo de una red neuronal.

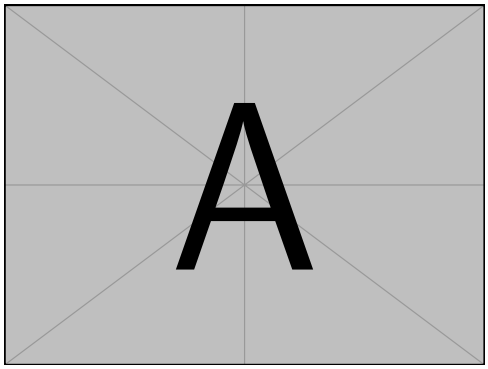
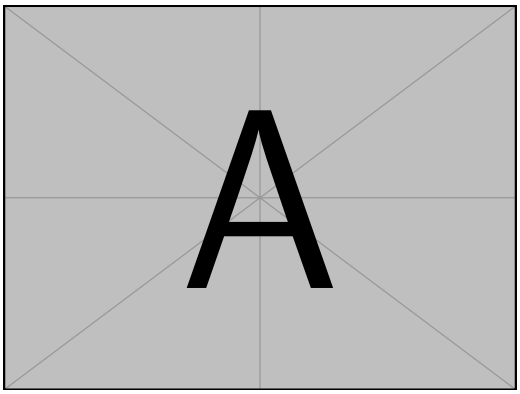


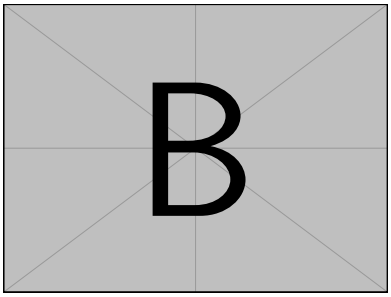
Figura 2.4: Ejemplo de como insertar figuras sencillas.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

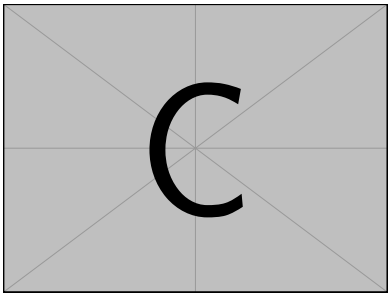
Las cantidades usan comas y punto decimal correctamente para México. Es decir 1,526.00, aparece correctamente tanto como texto, como ecuación $1,526.00$, cuando se agrega el comando `decimalpoint` en la configuración; test.



(a) Figure A



(b) Figure B



(c) Figure C

Figura 2.5: Using `subfigures` package.

2.3 Agregando Acrónimos y Glosario

A continuación se muestran algunos ejemplos del uso del glosario y los comandos que lo invocan. Por ejemplo, para hablar del lenguaje Latex y su especial aplicación en todo tipo de documentos que contienen ecuaciones matemáticas. Las Formulas que aparecen en los documentos son renderizadas de forma adecuada y fácil una vez que se acostumbra al uso de los comandos.

Por otro lado, dado un conjunto de resultados numéricos, existen diferentes métodos básicos para calcular el Error Cuadrático Medio, el cual se abrevia como ECM. Este error es común de usarse en en el cálculo de las estimaciones por Mínimos Cuadrados, *Least Squares* (LS).

2.4 Secciones de código

```
1  import numpy as np
2
3  def incmatrix(genl1,genl2):
4      m = len(genl1)
5      n = len(genl2)
6      M = None #to become the incidence matrix
7      VT = np.zeros((n*m,1), int)  #dummy variable
8
```

```
    #compute the bitwise xor matrix
    M1 = bitxormatrix(genl1)
    M2 = np.triu(bitxormatrix(genl2),1)

    for i in range(m-1):
        for j in range(i+1, m):
            [r,c] = np.where(M2 == M1[i,j])
            for k in range(len(r)):
                VT[(i)*n + r[k]] = 1;
                VT[(i)*n + c[k]] = 1;
                VT[(j)*n + r[k]] = 1;
                VT[(j)*n + c[k]] = 1;

                if M is None:
                    M = np.copy(VT)
                else:
                    M = np.concatenate((M, VT), 1)

                VT = np.zeros((n*m,1), int)

    return M
```

Apéndice A

Anexo o Apéndice

Contenido de los apéndices...

Bibliografía

[1] S. C. Chopra Abhimanyu, Prashar Abhinav, "Natural language processing," vol. 1, no. 4.

[2] A. C. Vásquez, H. V. Huerta, J. P. Quispe, and A. M. Huayna, "Procesamiento de lenguaje natural," vol. 6, no. 2, pp. 45–54.

[3] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," vol. 18, no. 5, pp. 544–551.

[4] S. E. M.-B. P. Bonet-Jover Alba, Sepúlveda-Torres Robiert, "Anotando la confiabilidad para mejorar la tarea de detecci on de desinformaci on: esquema de anotación, recurso y evaluación," vol. 70, pp. 15–26, 2023.

[5] V.-V. P. J. V.-G. R. Pan Ronghao, García-Días Jose Antonio, "Evaluación de modelos basados en transformadas para el sistema de recuperación de puntuación y mayúsculas en catalán y gallego," vol. 70, pp. 27–38, 2023.

[6] S. I. Giménez Raúl, "Ajuste y evaluación del modelo dialogpt sobre distintas colecciones de subtítulos de películas y series de televisión," vol. 70, pp. 63–71, 2023.

[7] M. L. Alarcón rodrigo, Martínez Paloma, "Ajuste de modelos bart para simplificación de textos sobre salud en español," vol. 70, pp. 111–122, 2023.

[8] J. López, "Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español," pp. 26–35, 2022.

[9] A. A. Ibáñez, "Desarrollo de modelos de deep learning para comprensión de textos usando técnicas NLP," 2017.

[10] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," vol. 32, no. 10, pp. 4291–4308. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[11] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and understanding neural models in NLP,"

[12] Y. Tsvetkov, M. Faruqui, W. Ling, G. Lample, and C. Dyer, "Evaluation of word vector representations by subspace alignment," pp. 2049–2054.

[13] A. Cano, "Aprendiendo vectores de palabras a partir de normas de asociación," 2020.

[14] P. A. M. Cáceres, A. C. Fuentes, I. C. Leiva, and E. J. Rojas, "Diseño y construcción de modelos de clasificación de incidentes de seguridad usando nlp en los registros de texto escrito para automatizar etiquetación," 2020.

[15] M. Jurado and H. Fellman, "Modelos de redes neuronales artificiales como sustento evaluativo al crecimiento pedagógico virtual en educación superior," 2020.

[16] A. Moreno, J. Díaz, L. Campillos, and T. Redondo, "Biomedical term extraction: Nlp techniques in computational medicine," 2018.

[17] N. Callaos, J. Horne, E. F. Ruiz-Ledesma, B. Sánchez, and A. Tremante, "Modelo con nlp y machine learning en la predicción de síntomas en base a conversaciones textuales con personas contagiadas de covid-19," 2023.

[18] B. R.-M. Barrera and D. C. Bárcena, "Análisis de informes médicos de pacientes mediante técnicas de nlp y consultas de metatesauros," 2021.

[19] B. Ayuga, "Diseño e implementación de un módulo para la estructuración de notas clínicas," 2018.

[20] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," vol. 20, no. 5, p. e10775.

[21] L. Zheng, Y. Wang, S. Hao, A. Y. Shin, B. Jin, A. D. Ngo, M. S. Jackson-Browne, D. J. Feller, T. Fu, K. Zhang, X. Zhou, C. Zhu, D. Dai, Y. Yu, G. Zheng, Y.-M. Li, D. B. McElhinney, D. S. Culver, S. T. Alfreds, F. Stearns, K. G. Sylvester, E. Widen, and X. B. Ling, "Web-based real-time case finding for the population health management of patients with diabetes mellitus: A prospective validation of the natural language processing–based algorithm with statewide electronic medical records," vol. 4, no. 4, p. e37.