Sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.

Pallo Tasiguano, Brandon Eduardo y Salazar Rivera, Adrian Alexander

Departamento de Ciencias de la Computación

Carrera de Ingeniería en Software.

Trabajo de titulación, previo a la obtención del título de Ingeniero en Software.

Uyaguari Uyaguari, Alvaro Danilo Msc.

Febrero del 2023

Latacunga

CERTIFICADO DE ANÁLISIS
magister

Tesina_Titulación_Pallo_Salazar_08-02-2023

8% Similitudes

< 1% Texto entre comillas
0% similitudes entre comillas

2% Idioma no reconocido

Nombre del documento: Tesina_Titulación_Pallo_Salazar_08-02-2023.docx
ID del documento: 6dccce48189acd58835b1f1d43e63b7f4d5c3b31
Tamaño del documento original: 673,28 ko

Depositante: JOSÉ LUIS CARRILLO
Fecha de depósito: 8/2/2023
Tipo de carga: interface
fecha de fin de análisis: 8/2/2023

Número de palabras: 9458
Número de caracteres: 64.263

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	Tesina_Titulación_Cruz_Quishpe_07-02-2023.docx | Tesina_Titulación_Cruz_... #ed7201 El documento proviene de mi biblioteca de referencias 2 Fuentes similares	3%		Palabras idénticas : 3% (251 palabras)
2	ebac.mx | Ciencia de datos: qué es y por qué es tan importante https://ebac.mx/blog/que-es-la-ciencia-de-datos	1%		Palabras idénticas : 1% (141 palabras)
3	dadun.unav.edu | DADUN: Análisis de las herramientas de procesamiento de lengu... https://dadun.unav.edu/handle/10171/60003	< 1%		Palabras idénticas : < 1% (77 palabras)
4	hdl.handle.net | Diskurtsoko koherentzia erlazioen iragarpenak euskararako BERT sa... http://hdl.handle.net/10810/58974	< 1%		Palabras idénticas : < 1% (31 palabras)
5	hdl.handle.net | Aplicación de la semántica multidimensional, alineamiento léxico-s... http://hdl.handle.net/10045/36401	< 1%		Palabras idénticas : < 1% (25 palabras)

Fuentes con similitudes fortuitas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	e-spacio.uned.es | Detección de eventos adversos en historiales clínicos mediante Pr... http://e-spacio.uned.es/fez/eserv/bibliuned:master-ETSInformatica-ICD-Dmartos/MartosLopez_Daniel_...	< 1%		Palabras idénticas : < 1% (32 palabras)
2	dspace.ups.edu.ec | Diseño de un sistema prototipo de diálogo persona-máquina ba... http://dspace.ups.edu.ec/bitstream/123456789/22403/1/UPS-CT009715.pdf	< 1%		Palabras idénticas : < 1% (33 palabras)
3	123dok.net | Reconocimiento de Entidades Biomédicas para el Español mediante la ... https://123dok.net/document/q7w77o0d-reconocimiento-entidades-biomedicas-combinacion-embeddi...	< 1%		Palabras idénticas : < 1% (29 palabras)
4	uvadoc.uva.es | Implementación de una herramienta basada en PLN para la detecci... https://uvadoc.uva.es/handle/10324/49995	< 1%		Palabras idénticas : < 1% (20 palabras)
5	doi.org | Natural language processing (NLP) tools in extracting biomedical concepts f... https://doi.org/10.1186/s12911-020-01352-2	< 1%		Palabras idénticas : < 1% (20 palabras)

Fuentes mencionadas (sin similitudes detectadas) Estas fuentes han sido citadas en el documento sin encontrar similitudes.

1 https://doi.org/10.1145/3366424.3382692
2 https://doi.org/10.1136/jamia.2009.002733
3 https://doi.org/10.1016/j.lpm.2021.102569
4 https://doi.org/10.1155/2021/6633213
5 https://doi.org/10.5772/51066

Firma:

.....

Uyaguari Uyaguari, Alvaro Danilo Msc.

C. C: 0103411112

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Departamento de Ciencias de la Computación

Carrera de Software

Certificación

Certifico que el trabajo de titulación: "**Sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.**" fue realizado por los señores **Pallo Tasiguano, Brandon Eduardo y Salazar Rivera, Adrian Alexander;** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Latacunga, 23 de febrero del 2023

Firma:

Uyaguari Uyaguari, Alvaro Danilo Msc.

C. C: 0103411112

Departamento de Ciencias de Computación

Carrera de Software

Responsabilidad de Autoría

Nosotros, **Pallo Tasiguano, Brandon Eduardo y Salazar Rivera, Adrian Alexander**, con cédulas de ciudadanía n° 1721332565 y n° 0550249056, declaro/declaramos que el contenido, ideas y criterios del trabajo de titulación**: "Sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.**" es de mi/nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, 23 de febrero del 2023

Firma

.....

Pallo Tasiguano, Brandon Eduardo

C. C: 1721332565

Firma

.....

Salazar Rivera, Adrian Alexander

C. C: 0550249056

Departamento de Ciencias de Computación

Carrera de Software

Autorización de Publicación

Nosotros **Pallo Tasiguano, Brandon Eduardo y Salazar Rivera, Adrian Alexander**, con cédula/cédulas de ciudadanía n° 1721332565 y n° 0550249056, autorizo/autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Título: "Sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.**" en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Latacunga, 23 de febrero del 2023

Firma

.....

Pallo Tasiguano, Brandon Eduardo

C. C: 1721332565

Firma

.....

Salazar Rivera, Adrian Alexander

C. C: 0550249056

DEDICATORIA

A mis queridos padres Eduardo y Patricia, quienes me han apoyado incondicionalmente en cada etapa de mi vida. Gracias por ser mis guía e inspiración, por enseñarme la importancia del esfuerzo y la perseverancia.

A mis hermanos Daniel, Josselyn y Gabriel, quienes han sido mis compañeros de vida y de aventuras. Gracias por estar siempre ahí para mí y por ser mi apoyo incondicional.

A mis abuelitos Luz y Manuel, quienes han sido mis guías y ejemplo de perseverancia y lucha. Gracias por sus sabios consejos y por su amor inmenso.

A mis primos Antony y Damarys, quienes han sido como mis hermanos. Gracias por estar siempre ahí para mí y por ser una parte importante de mi vida.

Y a mi tía Diana, quien siempre ha estado presente en mi vida y ha sido un gran apoyo emocional. Gracias por tu amor y tu dedicación.

Pallo Tasiguano, Brandon Eduardo

DEDICATORIA

A mi querido Dios, en primer lugar, por su amor y guía en mi vida. Por haberme permitido llegar a este momento tan especial de mi vida, la presentación de mi tesina.

A mis padres Alex y Herica por su constante apoyo y motivación. Gracias por haberme dado la educación y las herramientas necesarias para poder alcanzar mis metas. Su amor incondicional y dedicación son un verdadero ejemplo para mí.

A mis hermanos Evelyn y Mathias, por ser mi apoyo incondicional y por estar siempre presentes en los momentos más importantes de mi vida. A mi sobrina Emilia, por llenar de alegría mi vida y recordarme siempre la importancia de ver la vida con ojos de niño.

A mis primos Kevin y Daniel, gracias por su amistad y por haberme brindado su apoyo y consejos en momentos difíciles. A mi abuelito Gonzalo, gracias por ser un modelo de vida, por sus enseñanzas y por ser mi fuente de inspiración.

A mi tío Gonzalo Borja, ya que, desde el inicio de mi carrera, siempre has estado presente para brindarme tu sabiduría y consejos. Gracias por brindarme tu tiempo y dedicación, por motivarme a seguir adelante y por ayudarme a alcanzar mis metas. Tu orientación y consejos han sido invaluables para mi formación académica y personal.

Finalmente, a mi amiga Mery Arias, gracias por estar siempre a mi lado, por apoyarme en todo momento y por ser mi confidente en los momentos más difíciles.

A todos ustedes, les agradezco de todo corazón por haber formado parte de este camino que hoy culmina con la presentación de mi tesina. Sin su amor, apoyo y motivación, este logro no hubiera sido posible. Les agradezco por ser mi familia y amigos, los amo profundamente. Que Dios los bendiga siempre.

Salazar Rivera, Adrian Alexander

AGRADECIMIENTO

Deseo expresar mi más sincero agradecimiento a mis compañeros de universidad, quienes han sido parte fundamental en el proceso de desarrollo de esta tesina. Sus comentarios y sugerencias han sido invaluables para el éxito de este trabajo.

Asimismo, agradezco a mis docentes, quienes me brindaron los conocimientos y herramientas necesarias para llevar a cabo este proyecto. Su dedicación y compromiso con la enseñanza son dignos de admirar y agradezco por su tiempo y esfuerzo.

De manera especial, quiero agradecer a mi tutor, Msc Alvaro Uyaguari, por su orientación y consejos a lo largo de todo el proceso. Gracias por su tiempo, paciencia y compromiso en ayudar a alcanzar mis objetivos. Sus conocimientos y experiencia han sido una gran influencia en mi formación académica y personal.

Por último quiero agradecer a mi compañero de trabajo Adrián Salazar por su amistad, paciencia y colaboración en el desarrollo de nuestro trabajo de titulación.

Pallo Tasiguano, Brandon Eduardo

AGRADECIMIENTO

Quiero expresar mi más profundo agradecimiento por haberme acompañado en esta etapa tan importante de mi vida académica y personal.

Agradezco a Dios por su amor incondicional y por darme la fuerza y la sabiduría necesarias para lograr este logro.

A mis padres, por su amor y apoyo incondicional, por su confianza en mí y por ser mi fuente de inspiración. Gracias por su sacrificio y dedicación para que yo pudiera llegar hasta aquí.

A mis profesores, por su sabiduría y dedicación en la enseñanza, por su paciencia y su compromiso en mi formación académica. Gracias por compartir sus conocimientos y experiencias conmigo.

A mis compañeros, por su amistad y su compañía en este camino, por el apoyo mutuo y por hacernos crecer juntos. Gracias por ser una fuente de motivación y aprendizaje constante.

A mi tutor Msc Alvaro Uyaguari, por su orientación, paciencia y dedicación en mi formación académica. Gracias por ser un guía en este camino y por su apoyo incondicional.

A mi compañero de tesina Brandon Pallo, por su amistad, colaboración, paciencia y trabajo en equipo. Gracias por ser un compañero excepcional y por compartir conmigo esta experiencia tan enriquecedora.

Gracias a todos por ser parte de mi vida y por hacer posible este logro. Que Dios los bendiga siempre.

Salazar Rivera, Adrian Alexander

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

Actualmente algoritmos para el reconocimiento de entidades en el español son escasos, aún más si se trata del ámbito médico, es por eso que en este trabajo se desarrolló un sistema con el objetivo de reconocer y normalizar entidades biomédicas en español utilizando un corpus de reconocimiento de entidades dirigidas al dominio médico, particularmente en el caso de textos médicos en español, aplicando técnicas de Procesamiento de Lenguaje Natural (PLN) y el paradigma de cross-lingual (Usando recursos en el idioma inglés para trasladarlos al idioma español). La normalización de las entidades consiste en asignar una identificación única a cada entidad reconocida, lo que permite la integración de información de diferentes fuentes y la realización de análisis posteriores. Estos nuevos métodos y algoritmos serán integrados a un sistema web donde permita reconocer e identificar entidades médicas para posteriormente se pueda realizar algoritmos de predicción de diagnósticos a pacientes, aplicando herramientas, métodos y buenas prácticas de ingeniería de software. En resumen, el sistema desarrollado en esta tesina contribuye a mejorar la eficiencia y precisión en el reconocimiento y normalización de entidades biomédicas en español, lo que resulta de gran utilidad para la investigación y la práctica médica. La utilización de técnicas de cross-lingual y la integración de recursos en diferentes idiomas permiten ampliar el alcance y la precisión del sistema

Palabras clave: Procesamiento del lenguaje natural (PLN), Sistema web, entidades biomédicas

Abstract

Currently algorithms for the recognition of entities in Spanish are cases, even more if it is in the medical field, that is why in this work a system is developed with the objective of recognizing and normalizing biomedical entities in Spanish using a recognition corpus of entities directed to the medical domain, particularly in the case of medical texts in Spanish, applying Natural Language Processing (NLP) techniques and the cross-lingual paradigm (Using resources in the English language to translate them into the Spanish language). The normalization of the entities consists of assigning a unique identification to each recognized entity, which allows the integration of information from different sources and the performance of subsequent analysis. These new methods and algorithms will be integrated into a web system where they will make it possible to recognize and identify medical entities so that diagnostic prediction algorithms can later be made to patients, applying tools, methods and good software engineering practices. In summary, the system developed in this thesis contributes to improving the efficiency and precision in the recognition and normalization of biomedical entities in Spanish, which is very useful for research and medical practice. The use of cross-lingual techniques and the integration of resources in different languages allow to extend the scope and precision of the system.

Key words: Natural language processing (NLP), Web system, biomedical entities

Capítulo 1

Planteamiento del problema

En la actualidad las distintas técnicas de procesamiento de lenguaje natural y cross-lingual se han convertido en un campo muy importante de la inteligencia artificial, siendo así una herramienta muy importante para el desarrollo de distintos proyectos.

Existen algoritmos modernos utilizados para llevar a cabo procesos de predicción de diagnósticos y tratamientos de enfermedades, fundamentado en las notas médicas redactadas por el personal de salud durante una atención médica.

Sin embargo, para realizar el proceso de entrenamiento de dichos modelos es necesario identificar entidades en el texto como síntomas, diagnósticos, medicamentos, dolencias, entre otras, o así mismo apoyarnos en recursos lingüísticos existentes en español como corpus etiquetados, corpus paralelos, modelos pre entrenados y demás.

Los recursos lingüísticos en el dominio médico para identificar dichas entidades en el idioma español son pocos en comparación a otro idioma como el inglés, lo que limita la creación de nuevos modelos de aprendizaje automático en entornos médicos. Por esta razón proponemos realizar un sistema basado en la web para reconocer y normalizar entidades biomédicas, llevando a cabo técnicas de cross-lingual con el fin de incrementar los recursos biomédicos.

Basándonos en la problemática se formula la siguiente pregunta:

¿Cómo optimizar el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual?

Antecedentes

Actualmente la cantidad de datos médicos van aumentando exponencialmente, de lo que se estima que una parte de los mismos no se encuentran estructurados. Un solo paciente genera datos en exceso (historial médico, características, datos de pruebas, entre otros), si únicamente se pudiera aprovechar una pequeña parte de todos estos datos, la información obtenida sería muy valiosa. (Quevedo-Marcos, 2020)

El reconocimiento de entidades nombradas (Named Entity Recognition-NER) es una tarea importante en casi todas las áreas de aplicación del PLN como en la identificación y clasificación de información, para realizar dicha tarea existen algunas herramientas. No obstante, las herramientas más avanzadas tienen su aplicación en el reconocimiento de entidades como nombres, ciudades y fechas. (Castillo Molina et al., 2015)

El reconocimiento de entidades nombradas (NER) de texto biomédico es una de las tareas fundamentales de la minería de texto biomédico y tiene como objetivo identificar características especificadas como enfermedades procedentes del conjunto de texto biomédico. Los resultados de NER son a menudo otro objeto de la minería de texto. El texto biológico NER es la base de la investigación en bioinformática. (Ju et al., 2011)

La inteligencia artificial anexada al procesamiento de lenguaje natural (NLP), son los mejores procedimientos orientados a las mejoras en la atención médica (Jurafsky & Martin, 2008), ayudando a la mejora de resultados y para aumentar la eficiencia de la prestación de los servicios de salud. Sin embargo, la disponibilidad de recursos e instrumentos lingüísticos para el tratamiento correcto de textos distintos del español es insuficiente.

Justificación e importancia

El reconocimiento y la normalización de entidades biomédicas en notas clínicas, es un recurso fundamental para ayudar a los profesionales de la salud mediante encontrar rasgos en el texto a determinar unas posibles predicciones de diagnóstico y tratamientos médicos de un paciente. Existen varios enfoques y métodos para realizar este proceso de predicción, dependiendo de la naturaleza de los datos y de los recursos disponibles en el idioma en el que se redacten las notas médicas. (Quevedo-Marcos, 2020)

La escasez de información estructurada en el campo de la medicina, especialmente en el español a lo largo de los años, imposibilita la aplicación en esta área de nuevas tecnologías de Inteligencia Artificial relacionadas con el análisis de datos. Nuevas aplicaciones de PLN han sido creadas con el objetivo de procesar textos médicos de manera automática y aumentar así la cantidad de datos estructurados.

Para solventar las necesidades previamente descritas se decide realizar una investigación para desarrollar un sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual.

Objetivos

Objetivo General

Desarrollar un sistema web que permita el reconocimiento de entidades biomédicas mediante técnicas de cross-lingual utilizando modelos supervisados de inteligencia artificial.

Objetivos Específicos

- Explorar nuevos métodos para el reconocimiento y normalización de conceptos biomédicos.

- Desarrollar y aplicar nuevos métodos de reconocimiento de entidades biomédicas mediante el uso de técnicas de procesamiento del lenguaje natural con cross-lingual.
- Aplicar buenas prácticas en el ciclo de desarrollo e implementar el sistema con frameworks y arquitecturas actuales.

Variables de Investigación

Variable Independiente

Herramientas, modelos y métodos empleados en reconocimiento de entidades médicas.

Variable Dependiente

Desempeño del reconocimiento y normalización de entidades médicas.

Hipótesis

Si se construye el sistema de reconocimiento y normalización de entidades médicas que integre tantos modelos supervisados como etiquetadores automáticos, lograremos un desempeño similar o superior a los revisados en el estado del arte.

Capítulo 2

Marco teórico

En el presente capítulo se detallan definiciones, conceptos y la metodología que se seguirá para realizar el reconocimiento y normalización de entidades biomédicas. El mismo que se puede realizar mediante el uso de minería de datos que es una forma para poder identificar y analizar muchos tipos de entidades biomédicas (Habibi et al., 2017).

Otra importante herramienta para el reconocimiento y normalización de entidades biomédicas, es el NLP (Procesamiento de Lenguaje Natural), el cual se denomina encargado de reconocer, catalogar y organizar entidades de una manera efectiva para optimizar tiempo y recursos (Boudjellal et al., 2021). En el presente capítulo, se detallan las características más importantes de NLP y minería de datos, para posteriormente conocer los métodos y técnicas de reconocimiento de entidades biomédicas nombradas y los modelos secuenciales para el reconocimiento de entidades biomédicas.

Procesamiento de Lenguaje Natural (NLP)

Para el análisis, reconocimiento y normalización de entidades biomédicas, se ocupa como herramienta sistemas de extracción de información NLP los cuales ayudan a extraer conocimiento de datos textuales y no estructurados (Boudjellal et al., 2021).

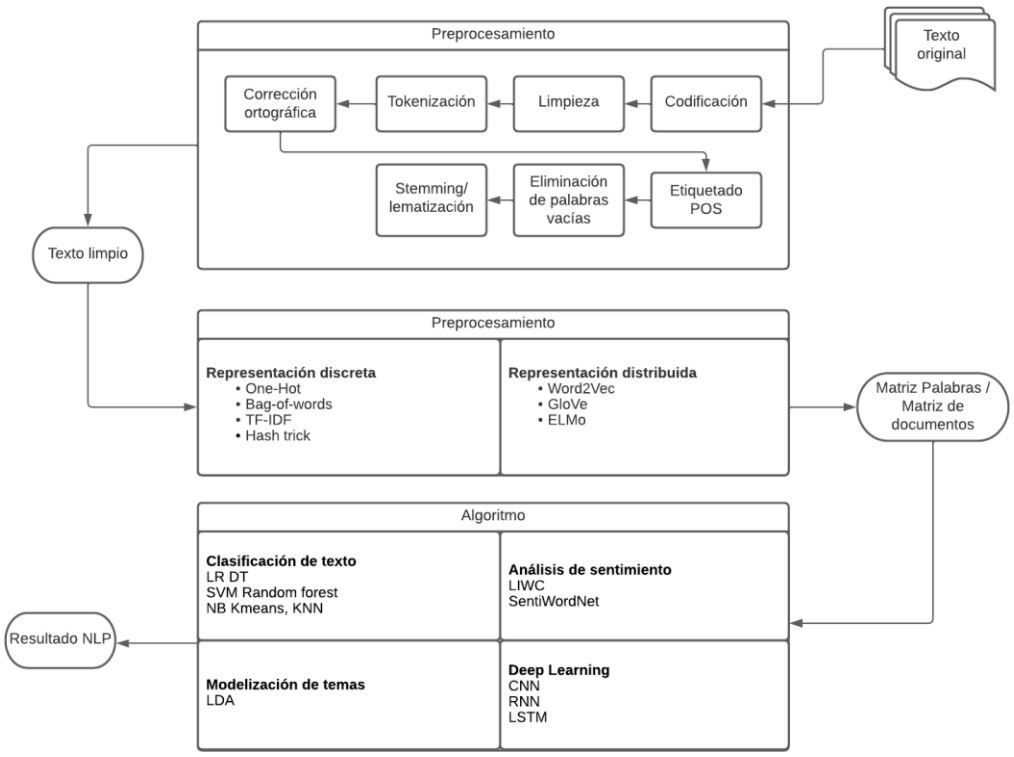
El procesamiento de lenguaje natural es aquel que abarca distintas metodologías y conceptos, con el objetivo de poder ser un medio de comunicación efectivo entre personas y ordenadores, de tal manera que el ordenador tenga un lenguaje mediante el cual pueda entender las órdenes que recibe de un humano (Kang et al., 2020). El procesamiento de lenguaje natural se subdivide en dos ramas importantes que son Comprensión de Lenguaje Natural (NLU), que se encarga de interpretar los información para obtener los datos más importantes y la segunda rama es la Generación de Lenguaje Humano (NLG) que a partir

de ciertos datos estructurados como fotos, video, texto o audio puede elaborar textos basados en lenguajes naturales, para el análisis humano (Kang et al., 2020).

Una tarea trascendental para que se utiliza NLP consiste en el reconocimiento de entidades con nombre (NER), que trata de delimitar y establecer tipos de categorías semantizadas (Tang et al., 2012). El proceso que ocuparemos de PNL consiste en tratamiento del texto, forma de representación del texto, preparación del modelo, validación del modelo (Kang et al., 2020).

Figura 1

Diagrama de flujo de NLP



Nota. Como observamos en la Figura 1 se muestra los procesos que se llevan a cabo el procesamiento de lenguaje natural, partiendo del pre procesamiento de texto para depurarlo de aquellos símbolos o signos, comprobar las faltas ortográficas, el etiquetado mediante un

tokenización POS, aminorar palabras mediante stemming. De la misma forma se debe escoger un tipo de sintaxis para identificar la información, para las tareas que cumplen los ordenadores al transformar palabras en vectores o matrices (Kang et al., 2020).

De manera que en los vectores de palabras se pueda aplicar algoritmos con el fin de tipificar, revisar sentimientos y obtener temas. Para posteriormente entrenar el modelo generado y entrar al proceso de evaluación de dicho modelo, para asegurar su calidad (Kang et al., 2020).

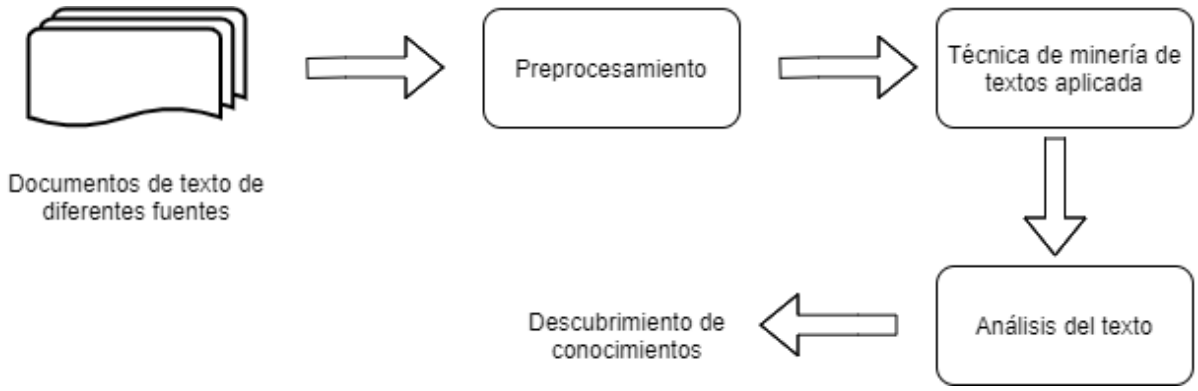
Minería de texto

Una manera más centrada para extraer conocimientos es la minería de textos, que nos ayuda al proceso de obtención de patrones importantes y descartables (Gaikwad et al., 2014). Lo que puede ser de mucha utilidad para el reconocimiento y normalización de entidades biomédicas.

La minería de textos tiene un proceso definido que arranca con un conjunto de documentos que vienen de múltiples fuentes, para procesarlos mediante un tipo de tratamiento poniendo énfasis en sus caracteres. Para proceder con el análisis del texto mediante el que se determina información de elevada precisión, es preciso nombrar que este proceso puede ser repetitivo dando como resultado una mejor calidad en la información. Con el resultante podemos generar conocimiento basado en la colección de documentos del incio (Gaikwad et al., 2014).

Figura 2

Diagrama de flujo de la Minería de Texto



Nota. La minería de textos se concentra en hallar patrones basándose en una base de datos extensa, con el fin de descubrir información importante y descartable basándose en textos que no tienen una estructura. Su proceso es ordenado como podemos observar en la Figura 2.

Métodos y técnicas de reconocimiento de entidades médicas nombradas

En este apartado abordaremos las técnicas y métodos más efectivos para el reconocimiento de entidades médicas, como lo son el uso de modelos con corpus etiquetados y con el uso de herramientas de etiquetado automático de entidades médicas.

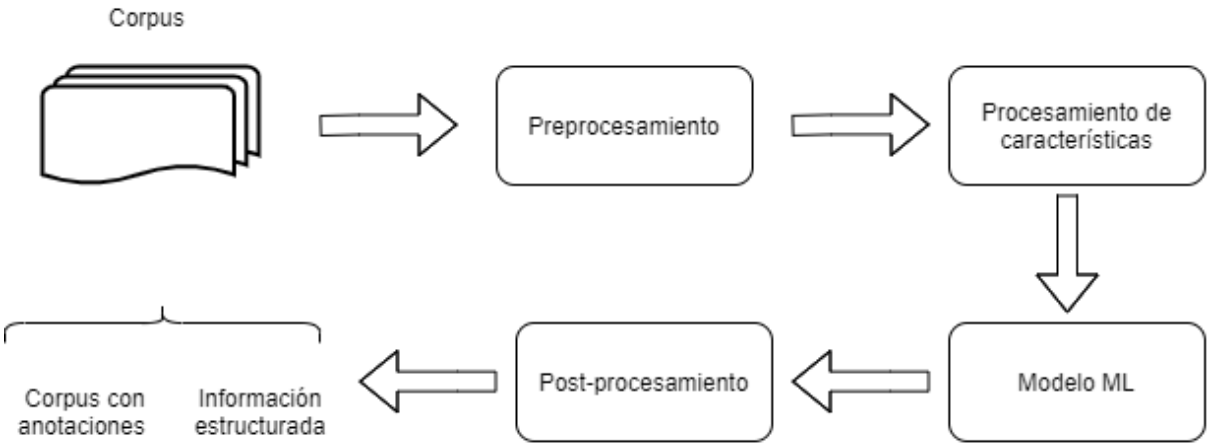
Es imprescindible recalcar que los modelos de machine learning que se entrenan basándose en conjuntos de datos etiquetados tienen capacidad de lograr mejores resultados en el ámbito de PNL clínico (Chen et al., 2015).

Usando modelos supervisados con corpus etiquetados

Los modelos supervisados con corpus etiquetados parten desde el corpus que se define como varios tipos de documentos en los que están embebidos anotaciones de uno o más clases de entidades. Las cuales son fundamentales para posteriormente servir de base para entrar un modelo basado en las anotaciones ante mencionadas (Campos et al., 2012).

Figura 3

Proceso general de los pasos necesarios para desarrollar soluciones NER basadas en ML



Nota. Para la organización del modelo supervisado con corpus etiquetados se sigue una estructura jerárquica que se detalla en la Figura 3, que comienza con el preprocesamiento de datos, procesar los rasgos más significativos, aplicación del modelo haciendo uso de las características producidas, procesamiento en donde se refinan las anotaciones para la obtención de nombres conocidos y finalmente el resultado un corpus correctamente procesado con información obtenida en un modelo organizado (Campos et al., 2012).

BIO Scheme

Para poder etiquetar un corpus, tenemos una variedad de formatos, en este caso vamos a analizar el BIO Scheme, se trata de un tipo de representación en la que se clasifica a la información mediante sufijos que son "B-" e "I-", y una clase "O" (Wu et al., 2017).

Utilizamos el prefijo "B-" cuando la palabra está en el inicio de la entidad con nombre, el prefijo "I-" significa que la palabra vigente se encuentra en la entidad con nombre, pero difiere de ser la primera palabra y por último "O" hace referencia a que la palabra no tiene ninguna relación con la entidad con nombre (Wu et al., 2017).

Modelos basados en redes tipo Transformer

BERT

A finales de 2018, los investigadores y trabajadores de Google diseñaron un nuevo modelo de PNL bautizado como BERT, sus siglas en el español significan: Representaciones de codificador bidireccional de transformadores. Es un modelo de lenguaje no supervisado profundamente bidireccional, siendo entrenado con anterioridad utilizando una enorme cantidad de un corpus de texto (Ayoub et al., 2021). Una de las desventajas que posee BERT es que es computacionalmente intensivo, lo que sería de gran complejidad la implementación del mismo sin recursos computacionales avanzados (Ayoub et al., 2021).

Aplicando un ejemplo para entender el contexto de BERT, se explica que el objetivo del PLN nos da una solución al predecir una palabra apoyándose en un entorno (Auquilla Vicuña & Mora Alvarez, 2022). De manera tradicional, esta predicción se realiza analizando una secuencia de texto durante un entrenamiento que sigue un flujo en una sola dirección: izquierda a derecha o una combinación derecha a izquierda, es por eso que es un codificador bidireccional de transformadores. (Auquilla Vicuña & Mora Alvarez, 2022). De esta forma es muy probable obtener resultados satisfactorios para aplicaciones o sistemas que requieran completar una palabra al final de una cadena, por ejemplo en la creación de frases.

Una vez entendido lo anterior decimos que BERT es un modelo que se entrena de manera bidireccional con el propósito de extraer un sentido más recóndito de un contexto (Auquilla Vicuña & Mora Alvarez, 2022). Para realizar lo anterior mencionado el modelo realiza un enmascaramiento de la palabra a predecir, extrayendo toda la información posible para determinar el contexto de la frase, tanto de izquierda a derecha como de derecha a izquierda, apoyándose en mecanismos que intervienen en el aprendizaje de las relaciones

entre palabras (Auquilla Vicuña & Mora Alvarez, 2022). Cabe destacar que previo al entrenamiento, el texto ingresado es convertido en una secuencia de tokens, a los que se les añaden ciertas etiquetas y metadatos (Auquilla Vicuña & Mora Alvarez, 2022).

RoBERTa

El modelo Transformer roBERTa es una variante del modelo Transformer original el mismo que fué entrenado en un corpus mucho más grande y con una serie de técnicas de optimización adicionales (Liu et al., 2019). El modelo roBERTa fue desarrollado por Facebook AI y se ha demostrado que supera a otros modelos de lenguaje populares en una serie de tareas de PLN (Liu et al., 2019).

Considerando que roBERTa fue entrenado en un corpus mucho más grande que el utilizado en el modelo BERT, nos permite una mayor comprensión del lenguaje (Liu et al., 2019). Además, se utilizaron técnicas de optimización complementarias, como la eliminación de las capas de atención para evitar el aprendizaje excesivo o en otras palabras el overfitting y el uso de una técnica de entrenamiento llamada "dynamic masking", para mejorar la capacidad del modelo para manejar palabras desconocidas. (Liu et al., 2019)

En términos de rendimiento, roBERTa ha logrado un rendimiento sobresaliente en una serie de tareas de PLN, incluyendo tareas de clasificación de sentimientos, extracción de entidades y resolución de anáforas (Liu et al., 2019). Por lo mencionado anteriormente, se ha demostrado que roBERTa es uno de los modelos de lenguaje más precisos y avanzados disponibles en la actualidad (Liu et al., 2019).

Usando herramientas de etiquetado automático de entidades médicas

Como un punto de partida para el reconocimiento y normalización de entidades médicas, se pueden utilizar distintas técnicas que pueden dar un diferente grado de efectividad. Para el

reconocimiento de entidades médicas existen herramientas que automatizan este proceso como lo son Metamap, Google NLP.

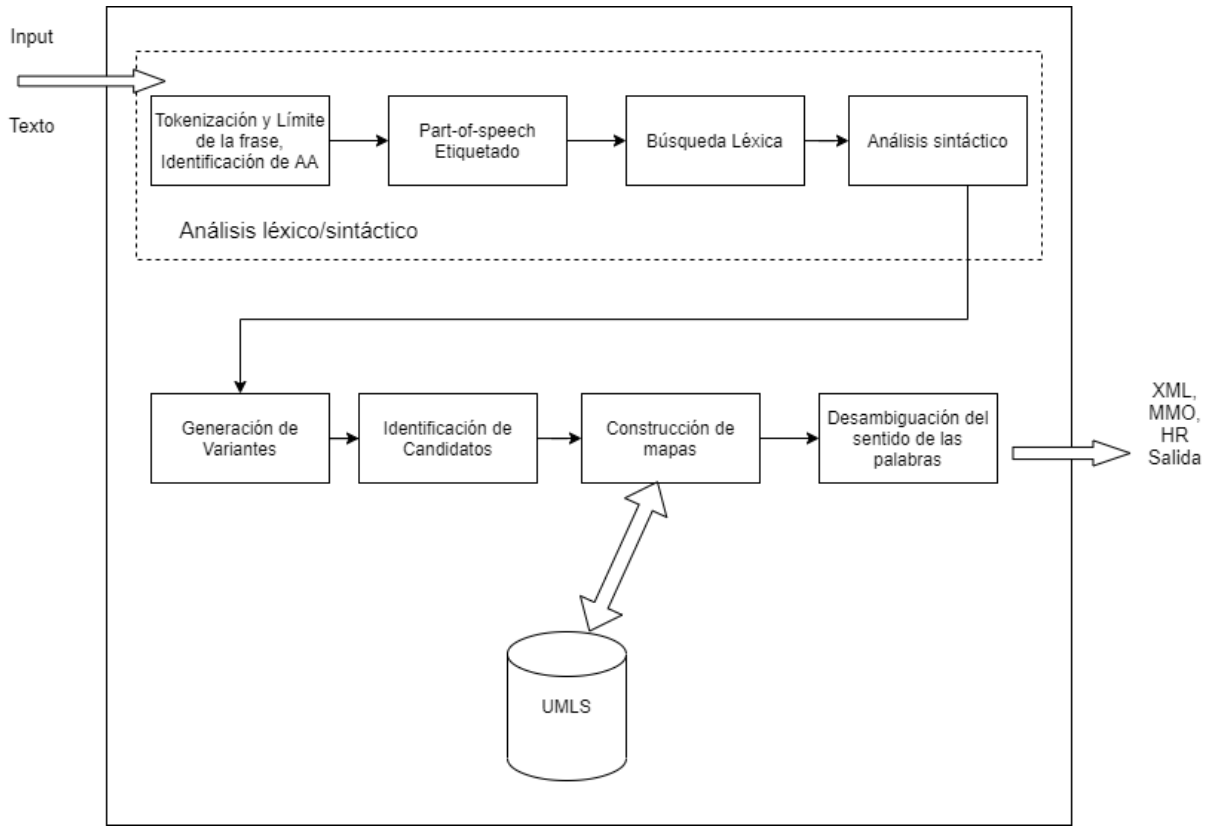
Que tienen aspectos que se relacionan, entre los cuales destacan que utilizan Unifed Medical Language System (UMLS), que hace referencia un diccionario biomédico completo y normalizado. El que utiliza metodologías de búsquedas en diccionarios y aprendizaje automatizado (Peng et al., 2020)

Metamap

Metamap es una herramienta de reconocimiento y extracción de información biomédica, fue desarrollada en 2001 por la Biblioteca Nacional de Medicina (Peng et al., 2020). Esta biblioteca de información biomédica nace como una necesidad de recuperar información, las fuentes que está herramienta utiliza es el sistema unificado de lenguaje médico (UMLS) (Aronson & Lang, 2010b).

Figura 4

Diagrama de procesos de la Herramienta Metamap.



Nota. En la figura 4 se detalla el proceso que maneja la herramienta Metamap, partiendo en el análisis sintáctico del texto, haciendo un preprocesamiento de palabras, para procesarlas y construir mapas en base a UMLS.

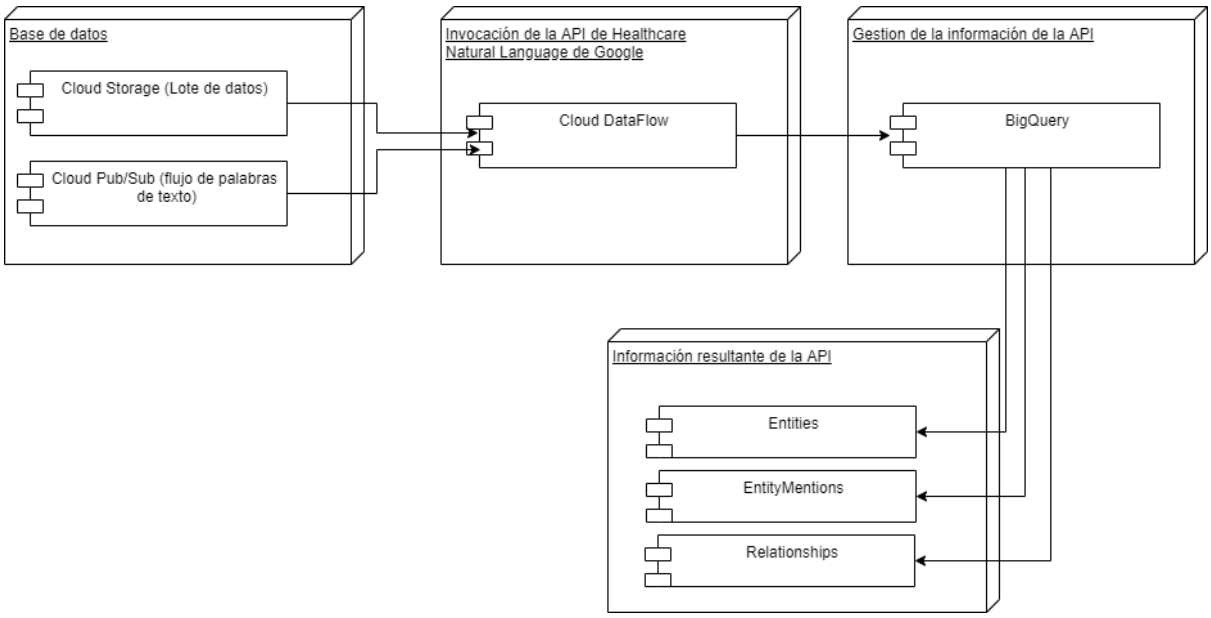
Google NLP

El API de Cloud Healthcare de Google nos sirve para poder realizar el reconocimiento y extracción de información biomédica, mediante el uso de técnicas para poder guardar, trasladar, interrelacionar datos del ámbito médico (*Descripción general de la API de Cloud Healthcare | API de Cloud Healthcare | Google Cloud*, s. f.). Esta herramienta hace énfasis en el análisis de textos, para hacer una representación de estos datos de manera ordenada, sus usos pueden ser la codificación clínica, extracción de entidades médicas y anotación de prescripciones técnicas.

De esta manera podemos obtener diferentes tipos de entidades relacionadas con el ámbito médico, tales como medicinas, tratamientos, enfermedades, aparatos médicos.

Figura 5

Diagrama de procesos de la herramienta Google NPL



Nota. En la Figura 5 se detalla el análisis que se hace desde recuperar datos desde la nube para posteriormente, analizarlos, organizarlos y obtener la información detallada en el último esquema de resultados de la figura 5.

Capítulo 3

Implementación del Sistema

En este capítulo se detallan todos los pasos que se efectuaron para desarrollar el sistema propuesto, el sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual. El funcionamiento de dicho sistema recibirá como entrada un texto, el mismo que se va a referir a un contexto médico y como salida obtendrá la extracción de datos, en este caso entidades médicas. Con el objetivo de tener una perpectiva de cómo funciona el sistema de reconocimiento de entidades, se explica de forma breve el proceso realizado: i) El primer paso es mediante un proceso NER para reconocer una entidad, puede ser una palabra o un conjunto de palabras que están referidas a una misma categoría, ii) Después de extraer la entidad, el siguiente paso es categorizar la entidad detectada. Las categorías de una entidad pueden ser de cualquier asunto, en este caso vamos a tener 4 tipo de entidades (NORMALIZABLES, NO_NORMALIZABLES, PROTEINAS y UNCLEAR).

Para el desarrollo del sistema, se utilizó una metodología de desarrollo de software ágil, que consiste en el trabajo en equipo, donde todos los miembros actúan juntos en entrega el proyecto en un tiempo y costo mínimo (Mahalakshmi & Sundararajan, 2013). La metodología utilizada para el desarrollo del sistema es Scrum, porque define el proceso de desarrollo del software como un conjunto de actividades flexibles que combinan herramientas y técnicas conocidas e implementables para que el equipo de desarrollo pueda implementar para construir el sistema. (Schwaber, 1997)

El proceso de trabajo de Scrum consiste en una estrecha colaboración del equipo scrum y Master con el Product Owner sobre iteraciones continuas del software en evolución, es decir que en scrum el proceso involucra un Scrum Master, el propietario del producto y

al equipo scrum, siendo el papel principal de los Scrum Masters el gestionar los procesos y eliminar los impedimentos que puedan afectar a la entrega del producto. (Srivastava et al., 2017) El equipo de scrum es multifuncional, ya que es conformado por desarrolladores, equipo de pruebas y otros expertos en diversos campos requeridos en el desarrollo del producto final, todo esto con el objetivo de cumplir con la satisfacción del cliente. (Srivastava et al., 2017)

Scrum proporciona una forma personalizada de trabajar, para manejar diferentes proyectos con distintos requisitos y tiene ventajas como la flexibilidad en la elección de los requisitos del sprint y la ausencia de procedimientos específicos a seguir. Para cumplir con los tiempos establecidos y evitar que distintos problemas perjudiquen al avance y desarrollo del producto se estableció dicha metodología mencionada en el texto, la cual trabaja con distintos términos que en la literatura (Srivastava et al., 2017), se menciona, lo mismo que fueron aplicados en el progreso de este proyecto y se explican a continuación

Sprint: es el periodo de tiempo más pequeño de scrum en el que el equipo trabaja para completar la tarea asignada, siendo la duración de cada uno de 1 a 3 semanas. El objetivo de cada sprint es entregar un producto potencialmente entregable. Al final de cada sprint hay una revisión del mismo que toma lugar con el propietario del producto para demostrar que el producto se puede enviar.

Sprint backlog: es una documentación, lista de tareas o requisitos identificados que se va a trabajar en el sprint actual.

Product backlog: es una lista de requisitos determinados por el propietario del producto y se llaman a las historias de usuario, seguido de la planificación de sprints.

Análisis y diseño del sistema

Una vez comprendido los componentes que involucran la metodología Scrum para la especificación de requisitos utilizaremos Historias de Usuario, de tal forma que podremos identificar los roles que se designará a cada uno de los miembros del equipo, involucrados en el desarrollo del proyecto, los cuales son: el Product Owner, siendo el responsable de definir los requisitos del producto, el seguimiento del proceso de desarrollo del producto e indica las prioridades de la lista de tareas, el Development Team es el equipo de desarrollo que se encargará de la ejecución de las tareas asignadas por el product owner y el Scrum Master que va a guiar y dirigir el desarrollo del producto aplicando la metodología. (Mariño & Alfonzo, 2014)

La distribución de roles de cada uno de los participantes en el proyecto se visualiza en la Tabla 1. Los roles fueron asignados por el scrum master, donde se muestra el rol, el integrante del equipo y la descripción de la función que va a desempeñar cada uno en el proyecto.

Tabla 1

Team Scrum

N°	Rol Scrum	Integrante	Funciones
1	Product Owner	Msc. Álvaro Uyaguari	Responsable de definir los requisitos del sistema, del seguimiento del proceso de desarrollo del producto y de indicar las

N°	Rol Scrum	Integrante	Funciones
			prioridades de la lista de tareas.
2	Scrum Master	Msc. Álvaro Uyaguari	Líder del equipo que va a guiar y dirigir el desarrollo del producto.
3	Development Team	Adrian Alexander Salazar Rivera Brandon Eduardo Pallo Tasiguano	Equipo de desarrollo que se encargará de la ejecución de las tareas asignadas para el desarrollo del sistema web para el reconocimiento y la normalización de entidades biomédicas.

Nota. En esta tabla se muestra la distribución de roles (rol, integrante y función) que va a desempeñar cada uno para el desarrollo del sistema web de reconocimiento y la normalización de entidades biomédicas en base a la metodología Scrum, debemos mencionar que debido a que el proyecto está conformado por dos integrantes, uno de ellos va a tomar el papel de Scrum Master que encima de realizar el trabajo correspondiente a su rol, también intervendrá en las actividades que pertenece al Development Team.

Una vez definido los roles y actividades para cada uno de los integrantes se lleva a cabo una reunión inicial liderado por el Scrum Master con el fin de adquirir la información necesaria para documentar las historias de usuario con ayuda del Product Owner y el Development Team.

Historias de usuario

Las historias de usuario se utilizan en metodologías ágiles para especificación de requisitos de forma rápida, sin tener que elaborar gran cantidad de documentos formales y sin requerir demasiado tiempo para administrarlos, además permiten responder rápidamente a los requisitos cambiantes. (Villamizar Suaza et al., 2015)

La Tabla 2 muestra las historias de usuario, en donde se detalla el nombre, el rol , la característica y la razón de la especificación de los requisitos del proyecto.

Tabla 2.

Historias de usuario

N° de historia de usuario	**Nombre**	**Rol**	**Característica**	**Razón**
1	H.U. 01	Como programador	Como programador quiero utilizar herramientas como metamap	Para el etiquetado automático de entidades médicas
2	H.U. 02	Como	Como programador	Para el etiquetado

N° de historia de usuario	Nombre	Rol	Característica	Razón
		programador	quiero utilizar herramientas como Google	automático de entidades médicas
3	H.U. 03	Como programador	Quiero crear un algoritmo que me permita tokenizar entidades.	Para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus
4	H.U. 04	Como programador	Quiero aplicar un modelo pre entrenado	Para reconocer entidades biomédicas a partir de un texto ingresado
5	H.U.06	Como programador	Quiero unificar mi algoritmo a las herramientas de metamap y Google	Para obtener un mejor reconocimiento de entidades biomédicas

Product Backlog del proyecto

Cuando las Historias de Usuario estén definidas la ejecución del product backlog es el siguiente paso, el cual consiste en una lista de requisitos determinados por el propietario del producto cuya finalidad es entregar el producto correctamente (Srivastava et al., 2017).

En la Tabla 3 se muestra el Product Backlog, en el cual están las historias de usuario que se desarrollarán en el transcurso del proyecto, con una estimación de tiempo en días, fecha de inicio, fecha final y el número de sprint perteneciente a cada historia de usuario.

Tabla 3.

Product Backlog

N° de historia de usuario	Nombre	Estimación (días)	Fecha de inicio	Fecha final	N° de Sprint
1	H.U. 01	10	24/10/2022	04/11/2022	1
2	H.U. 02	10	07/11/2022	18/11/2022	2
3	H.U. 03	15	21/11/2022	09/12/2022	3
4	H.U. 04	5	12/12/2022	16/12/2022	4
5	H.U. 05	20	19/12/2022	13/01/2023	6

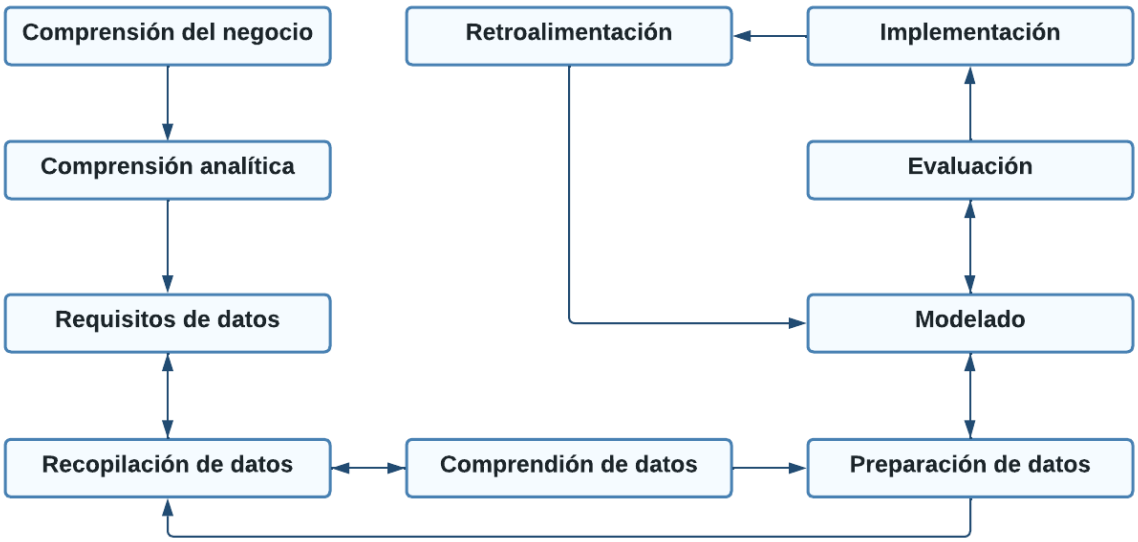
Metodología enfocada a la ciencia de datos

Además, para el desarrollo del sistema se empleó de manera breve una metodología enfocada en la ciencia de datos, la cual indica una rutina para encontrar soluciones a un problema en específico. Es un proceso cíclico que sufre un comportamiento crítico que guía a los analistas de negocios y científicos de datos a actuar en consecuencia.(Leiva, 2022)

Esquema del funcionamiento de la metodología enfocada a la ciencia de datos

Figura 6.

Metodología enfocada en la ciencia de datos



Nota. En la Figura 6 se representa la metodología aplicada, la cual consta de 10 etapas que forman un proceso iterativo para el uso de datos.

Comprensión del negocio: debemos tener la claridad de ¿cuál es el problema exacto que vamos a resolver?, en nuestro caso el reconocimiento de entidades en un contexto biomédico (Leiva, 2022).

Comprensión analítica: se determinan las técnicas de aprendizaje automático más aptas para la solución deseada, en nuestro caso técnicas de cross-lingual (Leiva, 2022).

Requisitos de datos: según los métodos seleccionados se decide qué contenido y formato deben tener los datos, en nuestro caso de un Corpus (Leiva, 2022).

Recopilación de datos: se reúnen los datos para estimar si son suficientes para solucionar el problema (Leiva, 2022).

Comprensión de datos: se analiza su contenido para determinar si hacen falta más datos (Leiva, 2022).

Preparación de datos: se preparan los datos de diversas fuentes si es necesario para que sean de mejor utilidad (Leiva, 2022).

Modelado: se usa la primera versión del grupo de datos para crear modelos predictivos o descriptivos, en nuestro caso para extraer entidades (Leiva, 2022).

Evaluación: se aplicarán varias pruebas para diagnosticar la efectividad del resultado que el modelo identificó en un principio (Leiva, 2022).

Implementación: al desarrollar y validar el modelo se implementa en nuestro sistema (Leiva, 2022).

Retroalimentación: se obtiene el feedback sobre el rendimiento del modelo implementado. Los científicos de datos lo utilizan para mejorar la precisión y utilidad del modelo (Leiva, 2022).

Diseño del sistema

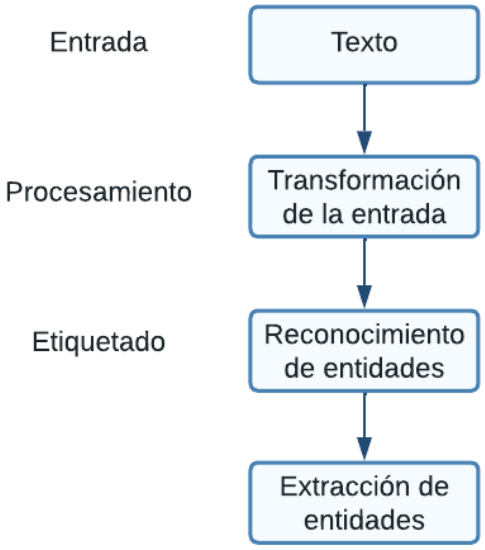
En esta sección se describe el diseño del sistema que se utiliza para el desarrollo del proyecto.

Esquemas del sistema de reconocimiento y la normalización de entidades biomédicas

A continuación, se presentan esquemas que representan el funcionamiento del sistema para el reconocimiento y la normalización de entidades biomédicas

Figura 7.

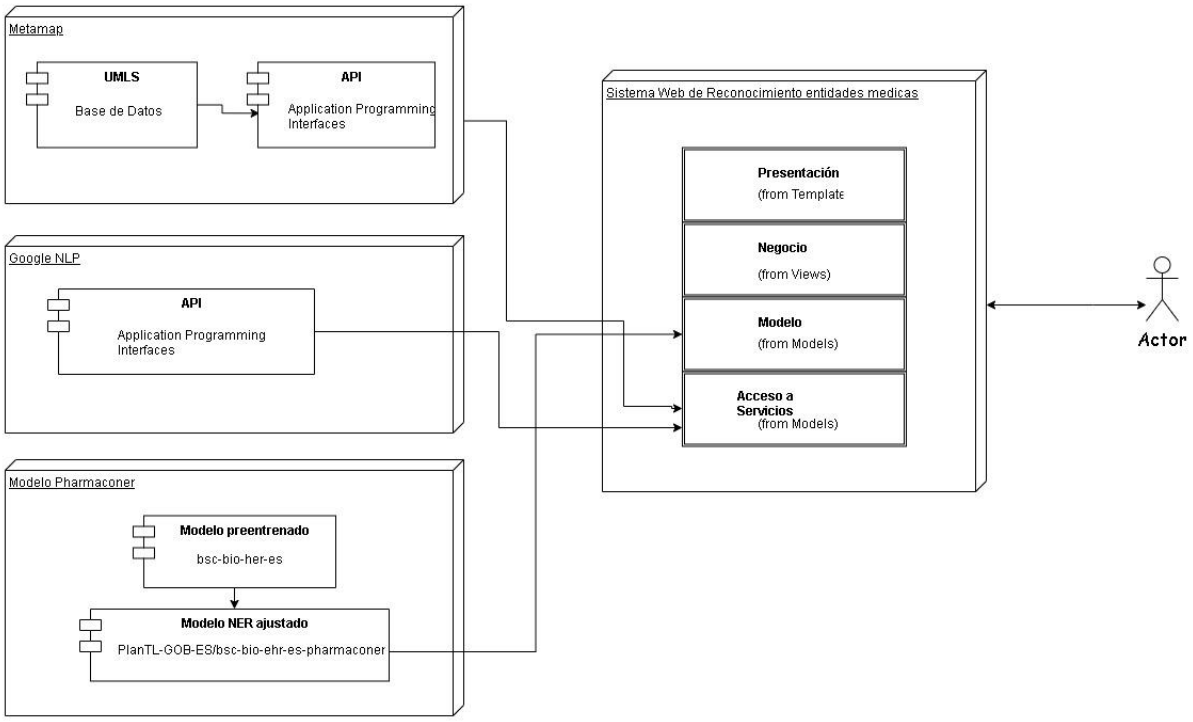
*Esquema del funcionamiento d*el sistema para el reconocimiento y la normalización de entidades biomédicas



Nota. En la Figura 7 muestra el funcionamiento del sistema primero, toma como entrada un texto en el idioma español, para posteriormente transformarlo en un conjunto de tokens, asignando a cada uno las características provenientes del corpus de donde se está extrayendo esta información. El siguiente paso es asignar a cada token una etiqueta de acuerdo a la entidad nombrada que le corresponde. Por último, se encarga de extraer las palabras que fueron etiquetadas como entidades nombradas y las muestra al usuario.

Figura 8

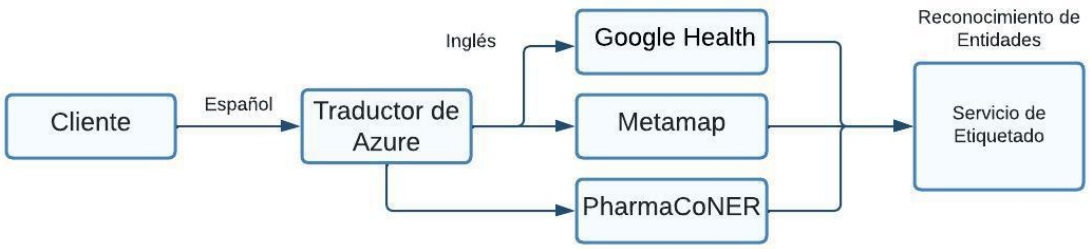
Diagrama para el reconocimiento y normalización de entidades biomédicas



Nota. El Sistema Web de Reconocimiento y normalización de entidades biomédicas utiliza el diagrama mostrado en la figura 8, en el cual se detalla la arquitectura usada, la cual contiene las herramientas de Metamap, Google NLP y el modelo PharmaCoNER.

Figura 9

Diagrama del sistema incorporando Metamap, Google NLP y el modelo entrenado sobre corpus.



Nota. El proceso esquematizado que seguirá el sistema planteado es el que se detalla en la Figura 9 en el cual abarcamos las herramientas descritas anteriormente en el marco teórico, de la misma forma el Corpus etiquetado de PharmaCoNER seleccionado para el reconocimiento de entidades biomédicas.

Herramientas empleadas para el desarrollo del sistema

Para el desarrollo del sistema, específicamente del modelo se eligieron diversas herramientas las cuales nos van a servir para llegar a nuestro objetivo, se explicarán posteriormente sin dejar de lado sus características, funcionamientos y otros componentes que se necesitan para su correcto funcionamiento.

Tabla 4

Herramientas utilizadas

Nombre	Descripción
Python	Lenguaje de programación utilizado para el desarrollo del sistema (Versión 3.8)
Visual Studio	Editor de código utilizado
Paper Space	Software que permite el uso de GPU en la nube
PharmaCoNER	Corpus de textos médicos especializados en farmacología y química farmacéutica
PyTorch	Biblioteca de aprendizaje profundo de código abierto muy popular y versátil que brinda a los desarrolladores una variedad de características y beneficios para desarrollar modelos sofisticados

Nombre	Descripción
	de aprendizaje profundo
Hugging Face	Es una de las plataformas más importantes en el campo del procesamiento del lenguaje natural

Nota. En la Tabla 4 se muestran las herramientas que utilizamos para el desarrollo del sistema.

Selección de API para generar el modelo (Pytorch y Hugging Face)

PyTorch es una potente y popular biblioteca de aprendizaje profundo de código abierto que ha experimentado un enorme crecimiento en popularidad en los últimos años. Permite a los desarrolladores crear rápida y fácilmente modelos complejos de aprendizaje profundo. PyTorch tiene varias características y beneficios que lo convierten en una opción popular para aplicaciones de aprendizaje profundo, tiene una interfaz fácil de usar que facilita el desarrollo de modelos, así como un fuerte soporte para GPU que permite tiempos de entrenamiento más rápidos (Min et al., 2021). Asimismo, PyTorch tiene una comunidad activa que brinda recursos y tutoriales útiles para ayudar a los desarrolladores a comprender y usar la biblioteca (Min et al., 2021). PyTorch tiene una amplia gama de herramientas y bibliotecas, lo que lo convierte en una excelente opción para proyectos complejos, convirtiéndose en una herramienta altamente extensible, lo que permite a los desarrolladores personalizarlo y ampliarlo para su caso de uso particular. (Min et al., 2021) En general, PyTorch es una biblioteca de aprendizaje profundo de código abierto muy popular y versátil que brinda a los desarrolladores una variedad de características y beneficios para desarrollar modelos sofisticados de aprendizaje profundo (Min et al., 2021).

La implementación de modelos de aprendizaje automático en producción puede ser un proceso desafiante, que requiere una cuidadosa consideración de una variedad de factores. Se destaca que la capacidad de iterar rápidamente en los experimentos es una gran ventaja de PyTorch y su flexibilidad para crear capas y arquitecturas personalizadas, sin embargo, esto puede crear desafíos en la producción, ya que la cantidad de modelos, parámetros y capas puede volverse difícil de administrar. (Argyriou et al., 2020)

En conclusión, PyTorch es un marco de aprendizaje profundo potente y fácil de usar que permite un desarrollo de aprendizaje profundo eficiente y efectivo. Permite a los desarrolladores crear, implementar y personalizar rápidamente modelos de aprendizaje profundo con facilidad y flexibilidad. Además, sus capacidades de capacitación distribuida y la integración con otros marcos populares como TensorFlow lo convierten en una herramienta ideal para aplicaciones de aprendizaje profundo. Con una interfaz de programación amigable y una variedad de herramientas de biblioteca, PyTorch es ideal tanto para principiantes como para expertos en aprendizaje profundo. El marco simple e intuitivo de PyTorch permite una experimentación y un desarrollo rápidos y efectivos, lo que la convierte en la principal herramienta para todas las necesidades de aprendizaje profundo.

La biblioteca Hugging Face es una de las plataformas más importantes en el campo del procesamiento del lenguaje natural (NLP, por sus siglas en inglés), desde el 2016, ha crecido rápidamente hasta convertirse en una de las bibliotecas de modelos de NLP más utilizadas y respetadas en la industria. (O’Neill, 2022)

Uno de los factores clave detrás del éxito de Hugging Face es su enfoque en la facilidad de uso. La biblioteca ha sido diseñada para ser intuitiva y accesible para desarrolladores de todos los niveles de habilidad, desde principiantes hasta expertos (O’Neill, 2022). Además, el código de la biblioteca está disponible en GitHub, lo que significa que los usuarios pueden contribuir y mejorar la biblioteca de manera constante.

Otro aspecto importante es la cantidad y calidad de los modelos disponibles en la biblioteca. Hugging Face ofrece una amplia gama de modelos pre-entrenados en diversas tareas de NLP, incluyendo análisis de sentimiento, traducción automática, extracción de información y generación de texto. (O'Neill, 2022) Estos modelos son el resultado de años de investigación y desarrollo por parte de los mejores equipos de NLP en todo el mundo, y están disponibles para su uso inmediato en la biblioteca.

Asimismo, la biblioteca también permite a los usuarios entrenar sus propios modelos basados en sus datos y configuraciones específicas. Esto es especialmente útil para las organizaciones que tienen datos confidenciales o que necesitan modelos altamente personalizados para sus aplicaciones.

Otro factor importante es la escalabilidad de la biblioteca. Hugging Face es compatible con una amplia gama de plataformas y sistemas operativos, y puede ser integrado fácilmente en aplicaciones más grandes (O'Neill, 2022). Además, la biblioteca es compatible con GPUs, lo que significa que los usuarios pueden entrenar modelos de forma rápida y eficiente.

En resumen, la biblioteca Hugging Face es una de las plataformas de NLP más importantes disponibles en la industria actual. Con su enfoque en la facilidad de uso, la calidad y cantidad de modelos disponibles, la posibilidad de entrenar modelos personalizados y la escalabilidad, Hugging Face se ha convertido en una herramienta esencial para cualquier persona o organización interesada en el procesamiento del lenguaje natural.

Selección de corpus

A continuación, detallaremos el corpus que nos será útil como recursos para hacer la validación del sistema. PharmaCoNER nos servirá para el proceso de identificación de entidades biomédicas.

PharmaCoNER

Un corpus lingüístico es un conjunto de documentos lingüísticos seleccionados y ordenados según criterios lingüísticos explícitos con el objetivo de ser usados como muestra del lenguaje («Corpus lingüístico», 2022). El corpus puede consistir en diferentes tipos de textos, como discursos, narrativas, diálogos, entrevistas, artículos, etc. Los textos del corpus pueden ser escritos o hablados, y se usan para estudiar el uso del lenguaje, las estructuras gramaticales, las relaciones entre palabras y frases, etc («Corpus lingüístico», 2022). Los corpus se pueden utilizar para diversos propósitos, como el análisis del lenguaje, la detección de patrones, el aprendizaje automático, la traducción automática, etc. («Corpus lingüístico», 2022)

Para el desarrollo de nuestro proyecto seleccionamos “PharmaCoNER” el mismo es un corpus de textos médicos especializados en farmacología y química farmacéutica (Gonzalez-Agirre et al., 2019). Este corpus es una valiosa herramienta para la investigación en el campo de la farmacología y la química farmacéutica, ya que proporciona una gran cantidad de textos de alta calidad y especializados en el área, que pueden ser utilizados para entrenar y evaluar modelos de procesamiento de lenguaje natural. (Gonzalez-Agirre et al., 2019)

El corpus PharmaCoNER es una colección de documentos que incluyen información sobre diferentes aspectos de la farmacología y la química farmacéutica, como la descripción de compuestos químicos, las interacciones farmacológicas y las propiedades

farmacocinéticas y farmacodinámicas de los medicamentos. (Gonzalez-Agirre et al., 2019) Estos documentos provienen de una amplia gama de fuentes, incluyendo artículos de revistas científicas, patentes, registros de ensayos clínicos y bases de datos de medicamentos (Gonzalez-Agirre et al., 2019).

El corpus PharmaCoNER es un recurso valioso para la investigación en farmacología y química farmacéutica debido a su gran cantidad de información especializada y su coherencia temática (Gonzalez-Agirre et al., 2019). Los investigadores pueden utilizar este corpus para entrenar modelos de procesamiento de lenguaje natural y desarrollar aplicaciones que extraigan información valiosa de textos médicos especializados (Gonzalez-Agirre et al., 2019). Por ejemplo en nuestro caso vamos a reconocer entidades biomédicas.

Además, el corpus PharmaCoNER es un recurso valioso para la formación de profesionales en el campo de la farmacología y la química farmacéutica, los estudiantes y profesionales pueden utilizar este corpus para familiarizarse con el lenguaje especializado utilizado en el área y mejorar su comprensión de los conceptos clave en farmacología y química farmacéutica. (Gonzalez-Agirre et al., 2019)

En conclusión, el corpus PharmaCoNER es un recurso valioso para la investigación y la formación en el campo de la farmacología y la química farmacéutica. Proporciona una gran cantidad de información especializada y coherente sobre diferentes aspectos de estas áreas, lo que nos va a servir para el desarrollo de nuestro proyecto.

Definición e implementación de modelos.

La metodología aplicada para el desarrollo del sistema nos señala que la siguiente etapa para el desarrollo del sistema es la planificación para cada sprint, priorizando las tareas más importantes de una forma ordenada, aplicando el Sprint Backlog. Para lograr lo mencionado

se llevaron a cabo reuniones presenciales y por videoconferencia en la plataforma Google Meet.

Sprint 01: Utilizar herramienta de metamap para etiquetado automático de entidades médicas

Para el proceso del Sprint 01, se tomó en consideración la Historia de Usuario (H.U.01), ubicada en la Tabla 2, en donde dice que se va a utilizar herramientas como metamap para el etiquetado automático de entidades médicas.

Historia de Usuario Detallada 1

Tabla 5

Historia de Usuario para utilizar herramientas como metamap para el etiquetado automático de entidades médicas.

Historia de Usuario	
Número: H.U.01	**Usuario:** Administrador
Nombre: Metamap para el etiquetado automático	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Medio
Duración: 10 días	**Interacción asignada:** 1
Responsables en el desarrollo: Adrian Salazar, Brandon Pallo	
Descripción: Como programador quiero utilizar herramientas como metamap para el etiquetado automático de entidades médicas.	

Validación:

- Se utilizó la herramienta de metamap para el etiquetado automático de entidades médicas.
- Se realizaron pruebas del funcionamiento del algoritmo.

Sprint Backlog 1

En la tabla 6 se muestra el sprint Backlog 01, donde se especifican las tareas realizadas en el sprint 1, también se muestra las horas empleadas en la actividad, fechas de inicio y fin, el responsable y el estado.

Tabla 6

Sprint Backlog 01

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.01	Selección y configuración de la herramienta	8	24/10/2022	24/10/2022	Adrian Salazar, Brandon Pallo	Finalizado
H.U.01	Instalar la imagen de metamap en doker	5	25/10/2022	26/10/2022	Adrian Salazar, Brandon Pallo	Finalizado
H.U.01	Aplicación y verificación	27	27/10/2022	04/11/2022	Adrian salazar, Brandon Pallo	Finalizado

de la
funcionalidad
de la
herramienta

Sprint 02: Utilizar herramienta de Google para etiquetado automático de entidades médicas

Para el proceso del Sprint 02, se tomó en consideración la Historia de Usuario (H.U.02), ubicada en la Tabla 2, en donde dice que se va a utilizar herramientas como Google para el etiquetado automático de entidades médicas.

Historia de Usuario Detallada 2

Tabla 7

Historia de Usuario para utilizar herramientas como Google para el etiquetado automático de entidades médicas.

Historia de Usuario	
Número: H.U.02	**Usuario:** Administrador
Nombre: Google para el etiquetado automático	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Medio
Duración: 10 días	**Interacción asignada:** 1
Responsables en el desarrollo: Adrian Salazar, Brandon Pallo	

Descripción: Como programador quiero utilizar herramientas como Google para el etiquetado automático de entidades médicas.

Validación:

- Se utilizó la herramienta de metamap para el etiquetado automático de entidades médicas.
- Se realizaron pruebas del funcionamiento del algoritmo.

Sprint Backlog 2

En la tabla 8 se muestra el sprint Backlog 02, donde se especifican las tareas realizadas en el sprint 2, también se muestra las horas empleadas en la actividad, fechas de inicio y fin, el responsable y el estado.

Tabla 8

Sprint Backlog 02

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.02	Selección y configuración de la herramienta	10	07/11/2022	09/11/2022	Adrian Salazar, Brandon Pallo	Finalizado
H.U.02	Aplicación de la herramienta	7	10/11/2022	11/11/2022	Adrian Salazar, Brandon Pallo	Finalizado
H.U.02	Verificación	23	14/11/2	18/11/2	Adrian Salazar,	Finalizado

de la funcionalidad de la herramienta	022	022	Brandon Pallo

Sprint 03: Crear un algoritmo que me permita tokenizar entidades para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus

Para el proceso del Sprint 03, se tomó en consideración la Historia de Usuario (H.U.03), ubicada en la Tabla 2, en donde dice que se va a crear un algoritmo que me permita tokenizar entidades para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus.

Historia de Usuario Detallada 3

Tabla 9.

*Historia de Usuario para c*rear un algoritmo que me permita tokenizar entidades para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus.

Historia de Usuario	
Número: H.U.03	**Usuario:** Administrador
Nombre: Desarrollo del algoritmo para la tokenización de entidades	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Medio
Duración: 15 días	**Interacción asignada:** 1

Historia de Usuario
Responsables en el desarrollo: Adrian Salazar, Brandon Pallo **Descripción:** Como programador quiero crear un algoritmo que me permita tokenizar entidades para reconocer entidades biomédicas a partir de textos médicos extraídos del corpus. **Validación:** - Se realizará el reconocimiento del corpus - Se realizará la construcción del algoritmo - Se realizará las pruebas del algoritmo

Sprint Backlog 3

En la tabla 10 se muestra el sprint Backlog 03, donde se especifican las tareas realizadas en el sprint 3, también se muestra las horas empleadas en la actividad, fechas de inicio y fin, el responsable y el estado.

Tabla 10.

Sprint Backlog 03

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.03	Indagación y experimentación del corpus	10	21/11/2022	23/11/2022	Adrián Salazar, Brandon Pallo	Finalizado
H.U.03	Construcción del algoritmo	25	24/11/2022	02/12/2022	Adrián Salazar, Brandon Pallo	Finalizado

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.03	Aplicación de pruebas al algoritmo	25	05/12/2022	09/12/2022	Adrián Salazar, Brandon Pallo	Finalizado

Sprint 04: Aplicar un modelo pre entrenado para reconocer entidades biomédicas a partir de un texto ingresado

Para el proceso del Sprint 04, se tomó en consideración la Historia de Usuario (H.U.04), ubicada en la Tabla 2, en donde dice que se va a aplicar un modelo pre entrenado para reconocer entidades biomédicas a partir de un texto ingresado.

Historia de Usuario Detallada 4

Tabla 11

Historia de Usuario para aplicar un modelo pre entrenado para reconocer entidades biomédicas a partir de un texto ingresado.

Historia de Usuario	
Número: H.U.04	**Usuario:** Administrador
Nombre: Aplicación de modelo pre entrenado para reconocer entidades biomédicas	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Medio
Duración: 5 días	**Interacción asignada:** 1

Historia de Usuario
Responsables en el desarrollo: Adrián Salazar, Brandon Pallo **Descripción:** Quiero aplicar un modelo pre entrenado para reconocer entidades biomédicas a partir de un texto ingresado. **Validación:** - Se aplicará modelo pre entrenado proveniente de la página hugging face. - Se realizará pruebas del modelo en plataformas como “paperspace” donde nos permitan el uso de GPU para su mejor rendimiento.

Sprint Backlog 4

En la tabla 12 se muestra el sprint Backlog 04, donde se especifican las tareas realizadas en el sprint 4, también se muestra las horas empleadas en la actividad, fechas de inicio y fin, el responsable y el estado.

Tabla 12.

Sprint Backlog 04

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.04	Aplicación del modelo en herramientas como paperspace donde nos	8	12/12/2022	13/12/2022	Adrián Salazar, Brandon Pallo	Finalizado

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
	permitan el uso de GPU					
H.U.04	Entrenamiento del modelo	5	14/12/2022	14/12/2022	Adrián Salazar, Brandon Pallo	Finalizado
H.U.04	Aplicación de pruebas para verificar la funcionalidad del modelo	7	15/12/2022	16/12/2022	Adrián Salazar, Brandon Pallo	Finalizado

Sprint 05: Unificar el algoritmo y el modelo pre entrenado para obtener un mejor reconocimiento de entidades biomédicas

Para el proceso del Sprint 05, se tomó en consideración la Historia de Usuario (H.U.05), ubicada en la Tabla 2, en donde dice que se va a unificar el algoritmo y el modelo pre entrenado para obtener un mejor reconocimiento de entidades biomédicas.

Historia de Usuario Detallada 5

Tabla 13

Historia de Usuario para unificar el algoritmo y el modelo pre entrenado para obtener un mejor reconocimiento de entidades biomédicas

Historia de Usuario	
Número: H.U.06	**Usuario:** Administrador
Nombre: Unión del algoritmo y el modelo pre entrenado	**Número de Sprint:** 1
Prioridad: Alta	**Riesgo de desarrollo:** Medio
Duración: 20 días	**Interacción asignada:** 1

Responsables en el desarrollo: Adrián Salazar, Brandon Pallo

Descripción: Unificar el algoritmo y el modelo pre entrenado para obtener un mejor reconocimiento de entidades biomédicas

Validación:

- Se integrarán las herramientas de Google y Metamap.
- Se desarrollará una interfaz aplicando un framework adecuado.
- Se integrará el modelo pre entrenado al sistema.
- Se realizará las pruebas correspondientes.

Sprint Backlog 5

En la tabla 14 se muestra el sprint Backlog 06, donde se especifican las tareas realizadas en el sprint 6, también se muestra las horas empleadas en la actividad, fechas de inicio y fin, el responsable y el estado.

Tabla 14.

Sprint Backlog 05

H.U.	TAREA	HORAS	INICIO	FIN	RESPONSABLE	ESTADO
H.U.05	Desarrollo de una interfaz web apoyándonos en el framework Django	15	19/12/2022	22/12/2022	Adrián Salazar, Brandon Pallo	Finalizado
H.U.05	Integración de las herramientas de metamap y google al sistema	20	23/12/2022	29/12/2022	Adrián Salazar, Brandon Pallo	Finalizado
H.U.05	Integración del código al sistema	10	30/12/2022	02/01/2023	Adrián Salazar, Brandon Pallo	Finalizado
H.U.05	Pruebas realizadas al sistema	35	03/01/2023	13/01/2023	Adrián Salazar, Brandon Pallo	Finalizado

Lenguaje de programación (Python), para realizar algoritmos y para la interfaz web

Algoritmo para el etiquetado de entidades médicas

Para el desarrollo de nuestro proyecto, utilizamos un algoritmo hecho a medida de los textos que vamos a clasificar, mediante el uso del lenguaje de programación Python. Como primer paso, clasificamos las posiciones que obtenemos de un archivo “.ann”, para posteriormente realizar el primer etiquetado de entidades del tipo (NORMALIZABLES, NO_NORMALIZABLES, PROTEINAS y UNCLEAR), con base en las posiciones antes obtenidas siguiendo las normas que se detallan en el BIO Scheme.

Con lo cual obtendremos un dataset listo con cada tipo de entidad etiquetado dependiendo de su relevancia en el contexto. Para utilizarlo en un modelo de Machine Learning

Figura 10

Algoritmo para la lectura de texto y clasificación de posiciones de las entidades

```python
#Leer el archivo de texto plano
def leerarchivotxt():
    txt=[]
    input_file_path = "S1139-76322017000200009-1.txt"
    with open(input_file_path, 'r') as f:
       txt=f.read()
    return txt

#Leer el archivo con las posiciones de las entidades
def leerarchivoann():
    input_file_path = "S1139-76322017000200009-1.ann"
    with open(input_file_path, 'r') as f:
        psc_proteinas = [line.split()
        for line in itertools.islice(f, 0, None,2)]
    return psc_proteinas

#Clasificar las posiciones y almacenarlas en arreglos
def clasificar_posc():
    psc_proteinas=leerarchivoann()
    proteinas = []
    proteinas_aux=[]
    for i in range(len(psc_proteinas)):
        psc = psc_proteinas[i]
        for j in range(1):
            proteinas_aux.append(psc[2])
            proteinas_aux.append(psc[3])
        proteinas.append(proteinas_aux)
        proteinas_aux=[]
    return proteinas
```

Figura 11

Algoritmo para el etiquetado de entidades siguiendo la norma BIO Scheme.

```
def tokenizar_bei():
  proteina=pln_text_corpus_BEI()
  bio_tags=[]
  proteinas=[]
  for i in range(len(proteina)):
    words = proteina[i].split()
    if(len(words)>1):
      for j in range(len(words)):
        words = proteina[i].split()
        words = words[j].split()
        proteinas.append(words)
        if(j==0):
         bio_tags.append("B")
        if(j==1):
          bio_tags.append("I")
        elif(j>=2):
          bio_tags.append("I")
    else:
      bio_tags.append("B")
      proteinas.append(words)
  for o in range(len(proteinas)):
    print(proteinas[o],bio_tags[o])
```

Algoritmo para la interfaz web

Para la interfaz web ocupamos el framework de desarrollo Django que trabaja con una frontend basado en HTML, Css y javascript. Además de un backend el que estará desarrollado usando el lenguaje de programación Python.

Figura 12

Algoritmo para la interfaz web.

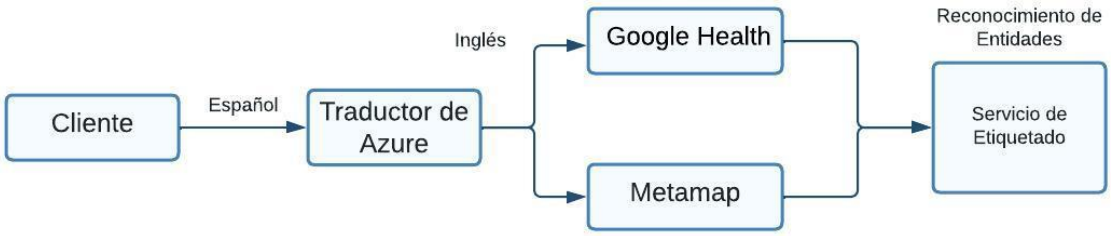
```
46      <div class="row">
47        <div class="col-sm">
48          <form method='GET' action="">
49            {% csrf_token %}
50            <label for="exampleFormControlTextarea1" class="form-label" aria-placeholder="Textarea'
51              su
52              frase</label>
53            <textarea class="form-control" id="exampleFormControlTextarea1" name="dato"
54              rows="4">{{cadena}}</textarea>
55              {% if messages %}
56  <ul class="messages">
57      {% for message in messages %}
58      <li{% if message.tags %} class="{{ message.tags }}"{% endif %}>
59        <script>window.alert("{{ message }}")</script></li>
60      {% endfor %}
```

Reconocimiento de entidades con técnicas de cross lingual (Traduciendo del español al inglés y etiquetando con Metamap y Google Health)

A Continuación, se muestra el reconocimiento de entidades con técnicas de cross lingual con las herramientas de Metamap y Google Health

Figura 13

Diagrama del reconocimiento de entidades con técnicas de cross lingual



Nota. En la Figura 13 se muestra el funcionamiento del reconocimiento de entidades biomédicas aplicando técnicas de cross lingual.

Usando modelos entrenados sobre corpus etiquetado.

Se utilizó un Modelo de lenguaje biomédico y clínico para el español pre entrenado a partir del corpus de PharmaCoNER. Específicamente el modelo es:

- bsc-bio-her-es: modelo entrenado en un corpus biomédico-clínico en español (*PlanTL-GOB-ES/bsc-bio-ehr-es · Hugging Face*, s. f.) .

Para la tarea de reconocimiento de entidades nombradas (NER) se encuentra ajustado el modelo pre entrenado y se lo encuentra en HuggingFace. Es el siguiente:

- bsc-bio-her-es-pharmaconer: modelo NER para sustancias, compuestos y proteínas en casos clínicos (*PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer · Hugging Face*, 2022).

Capítulo 4

Validación del Sistema

Las entidades biomédicas son una parte esencial del texto médico, por lo que es importante contar con un reconocimiento casi preciso y automatizado de estas entidades. En este capítulo abordaremos las actividades realizadas durante el periodo de este proyecto para obtener un mejor resultado en el reconocimiento de entidades con respecto al anterior trabajo realizado.

Para la validación del sistema recibirá como entrada un texto, el mismo que se va a referir a un contexto médico y como salida obtendrá la extracción de datos, en este caso entidades médicas. Con el fin de tener una visión sobre cómo funciona el sistema de reconocimiento de entidades, se explica de forma breve el proceso realizado: i) El primer paso es mediante un proceso NER detectar una entidad, puede ser una palabra o un conjunto de palabras que están referidas a una misma categoría, ii) Después de extraer la entidad, el siguiente paso es categorizar la entidad detectada. Las categorías de una entidad pueden ser de cualquier asunto, en este caso vamos a tener 4 tipo de entidades (NORMALIZABLES, NO_NORMALIZABLES, PROTEINAS y UNCLEAR).

Selección de las herramientas y recursos para la validación del sistema

Para la validación del sistema se eligieron diversas herramientas las cuales nos van a servir para llegar a nuestro objetivo, se explicarán posteriormente sin dejar de lado sus características, funcionamientos y otros componentes que se necesitan para su correcto funcionamiento.

Medline (Del Clef Gold Corpus)

Medline es una fuente de información creada por la Librería Nacional de Medicina (NLM), correspondiente al Clef Gold Corpus, es una colección de varios temas entre ellos biomedicina, ciencias de la vida, salud, ciencias químicas y bioingeniería entre otros. Temáticas imprescindibles en el conocimiento de profesionales de salud, científicos en investigación de ciencias biomédicas. (Costas et al., 2008).

EMEA (Del Clef Gold Corpus)

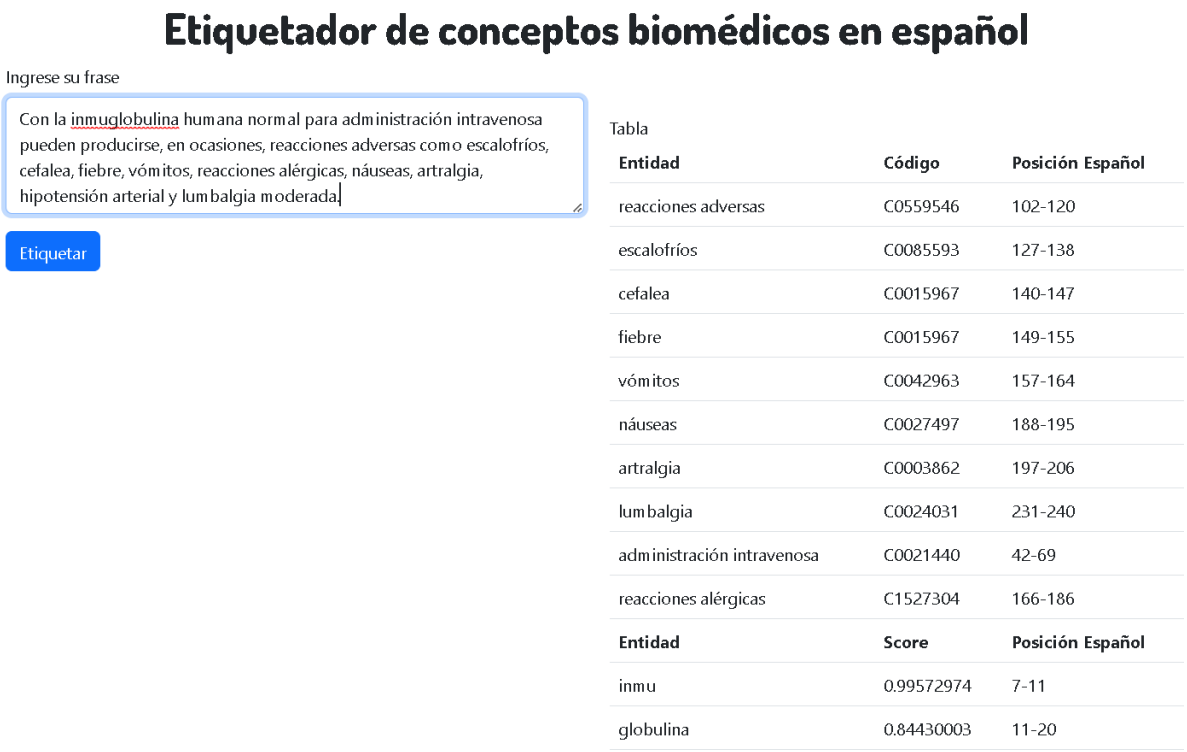
EMEA es un corpus de información biomédica correspondiente al Clef Gold Corpus el cual fue creado por la Agencia Europea de Medicina (EMEA). En este corpus se anexa información de medicamentos específicamente traducidos a varios idiomas de toda la Unión Europea. (CORRALES et al., s. f.)

Análisis de resultados

El sistema web para el reconocimiento y la normalización de entidades biomédicas mediante técnicas de cross-lingual ha sido evaluado y se presentan a continuación los resultados del análisis de los datos obtenidos.

Figura 14

Interfaz de usuario del Etiquetador de conceptos biomédicos



Nota. En la figura 14 se observa el sistema de etiquetado de conceptos biomédicos en español, donde se integró las herramientas de Metamap, Google NLP y por último el Corpus etiquetado de PharmaCoNER, donde se puede evidenciar que gracias a la integración de los mismos podemos obtener una mayor cantidad de entidades.

Efectividad de las herramientas de etiquetado

Tabla 15

Resultados de efectividad de las herramientas de etiquetado (Google NLP y Metamap)

Corpus	Etiquetado con combinación de herramientas (Google NLP y Metamap)	% de efectividad
Medline	166/316	52,53%
EMEA	237/430	55,11%

Nota. En la Tabla 15 se evidencia los resultados obtenidos al evaluar las herramientas de Google NLP y Metamap en base a los corpus de Medline y EMEA, donde se evidencia que evaluando el corpus de Medline se obtiene un resultado del 52,53%, ya que dichas herramientas reconocieron 166 entidades de 316 que fue su total. Por otra parte, al experimentar con el corpus de EMEA se obtiene un valor del 55,11%, puesto que por la combinación de estas dos herramientas se reconocieron 237 entidades de 430.

Tabla 16

Reconocimiento de nuevas entidades con PharmaCoNER

Corpus	Etiquetado con combinación de herramientas (Google NLP y Metamap)	Nuevas entidades identificadas con PharmaCoNER
Medline	166/323	7/323
EMEA	237/449	17/449

Nota. En la Tabla 16 se presenta el resultado de la integración de PharmaCoNER, el mismo que muestra que en el corpus de Medline reconoció 7 entidades nuevas y en el de EMEA 17 entidades nuevas, cabe recalcar que estas nuevas entidades las dos herramientas anteriores no pudieron identificar.

Tabla 17

Resultados de efectividad de las herramientas de etiquetado (Google NLP, Metamap y PharmaCoNER)

Corpus	Etiquetado con combinación de herramientas (Google NLP, Metamap y PharmaCoNER)	% de efectividad	% que incrementa al agregar PharmaCoNER con respecto a las anteriores dos herramientas
Medline	173/323	53,56%	1,03%
EMEA	254/449	56,57%	1,46%

Nota. En la Tabla 17 se evidencia los resultados obtenidos al evaluar las herramientas de Google NLP, Metamap y PharmaCoNER en base a los corpus de Medline y EMEA, donde se evidencia que evaluando el corpus de Medline reconoce 173 entidades de un total de 323, obteniendo un resultado del 53,56%. Cabe mencionar que gracias a la integración de esta última herramienta el porcentaje de efectividad aumentó en un 1,03%. Por otra parte en el corpus EMEA se reconocieron 254 entidades de 449, obteniendo un resultado de efectividad del 56,57%, destacando que la efectividad aumentó en un 1,46% en comparación a lo anterior.

Estos resultados se obtuvieron por la integración de estas tres herramientas de etiquetadores automáticos (Google NLP, Metamap y PharmaCoNER) dando como resultado una mejor etiquetación de entidades biomédicas, ya que cada herramienta obtiene diferentes etiquetas, como podemos evidenciar a continuación.

Tabla 18

Ejemplo de combinación de herramientas en el corpus Medline

Frase	Entidad	Tipo	Google NLP	Metamap	PharmaCoNER
Modificaciones de los valores de K + en la inducción con pentotal y succinilcolina.	K	CHEM	0	0	1
	succinilcolina	CHEM	1	1	0
	pentotal	CHEM	0	0	0

Nota. En la Tabla 18 presenta un ejemplo de cómo actúa la combinación de las herramientas donde el “1” representa que se encontró la entidad y “0” significa que no se encontró. Además, se evidencia a qué tipo semántico pertenece la entidad.

Tabla 19

Ejemplo de combinación de herramientas en el corpus EMEA

Frase	Entidad	Tipo	Google NLP	Metamap	PharmaCoNER
Retacrit fue tan eficaz como EPREX/ ERYPO para corregir y mantener los recuentos de glóbulos rojos.	EPREX	CHEM	0	0	1
	recuentos de glóbulos rojos	PHEN	0	1	0
	recuentos de glóbulos rojos	PROC	1	1	0
	Retracrit	CHEM	0	0	1
	ERYPO	CHEM	0	0	1

Nota. En la Tabla 19 presenta un ejemplo de cómo actúa la combinación de las herramientas donde el “1” representa que se encontró la entidad y “0” significa que no se encontró. Además, se evidencia a qué tipo semántico pertenece la entidad.

Herramienta web para la visualización de resultados

Para la visualización de resultados optamos por el uso de “Django", ya que es un framework para el desarrollo web de código abierto basado en Python, el mismo fué desarrollado originalmente para la gestión de contenido de un sitio web y ha evolucionado hasta convertirse en una plataforma poderosa y versátil para la creación de aplicaciones web. (Vidal-Silva et al., 2021)

Una de las principales características de Django es su enfoque en la reutilización de código, tiene una arquitectura modular que permite a los desarrolladores crear aplicaciones utilizando módulos y aplicaciones de terceros, esto significa que los desarrolladores pueden concentrarse en resolver problemas específicos y no tener que escribir todo el código desde cero. (Vidal-Silva et al., 2021) Además, Django ofrece una amplia gama de herramientas y paquetes para solucionar problemas comunes en la creación de aplicaciones web, como la gestión de formularios, la autenticación de usuarios y la generación de informes. (Vidal-Silva et al., 2021)

Otro aspecto importante de Django es su enfoque en la seguridad, ya que incluye una serie de medidas de seguridad integradas que protegen las aplicaciones de posibles ataques, como la inyección de SQL y el robo de sesión (Vidal-Silva et al., 2021).

Django también es conocido por su facilidad de uso, ofrece una interfaz administrativa intuitiva que permite a los desarrolladores y los usuarios administrar fácilmente la aplicación, utiliza un lenguaje de plantilla simple y eficiente que permite a los desarrolladores separar el

código HTML del código Python, lo que hace que el desarrollo de aplicaciones sea más fácil y mantenible, es por eso que hemos optado por el uso de este framework.

Capítulo 5

Conclusiones y Recomendaciones

Conclusiones

- El desarrollo del marco teórico permitió la obtención de conocimientos acerca de Procesamiento de Lenguaje Natural (NLP), Minería de texto, Métodos y técnicas de reconocimiento de entidades médicas nombradas, corpus etiquetados, BIO Scheme y las redes transformer BERT y RoBERTa
- El algoritmo implementado para etiquetar las entidades biomédicas basado en el corpus Pharmacomer, obtuvo un buen resultado, aumentando en 1,03% en la evaluación del corpus de Medline y un 1,46% en el corpus de EMEA como parte de la base del proceso de reconocimiento y normalización de entidades biomédicas.
- El uso de herramientas como paperspace fue de gran ayuda en el desarrollo del trabajo, ya que la misma poseía GPU en la nube brindándonos de la misma manera un mejor rendimiento con respecto a la GPU de nuestras máquinas.
- El modelo implementado ayudó al reconocimiento de nuevas entidades, en su mayoría del tipo CHEM (Chemicals & Drugs).

Recomendaciones

- Es importante revisar la literatura existente sobre técnicas de reconocimiento y normalización de entidades biomédicas, así como también sobre técnicas de cross-lingual. Esto permitirá identificar las técnicas más avanzadas y las limitaciones actuales.
- Es fundamental definir claramente el problema que se desea abordar, en términos de los tipos de entidades biomédicas que se desean reconocer y normalizar, los idiomas involucrados y las limitaciones de los datos disponibles.
- Es importante seleccionar conjuntos de datos adecuados para el entrenamiento y evaluación del sistema. Estos datos deben representar la diversidad de entidades biomédicas y de idiomas involucrados.
- Se deben seleccionar las técnicas más adecuadas para el reconocimiento y la normalización de entidades biomédicas, considerando tanto técnicas de procesamiento de lenguaje natural como técnicas de cross-lingual. Es importante evaluar y comparar diferentes enfoques.
- Se debe implementar un sistema web que permita el reconocimiento y la normalización de entidades biomédicas en diferentes idiomas. Este sistema debe ser fácil de usar y permitir la interacción con los usuarios.
- Se debe realizar una evaluación exhaustiva del sistema, en términos de su precisión y cobertura. Es importante realizar comparaciones con otros sistemas existentes y establecer las limitaciones y áreas de mejora.
- Se deben analizar los resultados obtenidos y sacar conclusiones sobre la efectividad del sistema. Se deben discutir las limitaciones del sistema y las posibles áreas de mejora para futuras investigaciones.

Bibliografía

Argyriou, A., González-Fierro, M., & Zhang, L. (2020). Microsoft Recommenders: Best Practices for Production-Ready Recommendation Systems. *Companion Proceedings of the Web Conference 2020*, 50-51. https://doi.org/10.1145/3366424.3382692

Aronson, A. R., & Lang, F.-M. (2010a). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236.

Aronson, A. R., & Lang, F.-M. (2010b). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229-236. https://doi.org/10.1136/jamia.2009.002733

Auquilla Vicuña, J. F., & Mora Alvarez, J. C. (2022). *Diseño de un sistema prototipo de diálogo persona-máquina basado en la arquitectura BERT* [BachelorThesis]. http://dspace.ups.edu.ec/handle/123456789/22403

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management*, *58*(4), 102569. https://doi.org/10.1016/j.ipm.2021.102569

Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., & Dai, L. (2021). ABioNER: A BERT-Based Model for Arabic Biomedical Named-Entity Recognition. *Complexity*, *2021*, 6633213. https://doi.org/10.1155/2021/6633213

Campos, D., Matos, S., & Oliveira, J. L. (2012). Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools. En S. Sakurai (Ed.), *Theory and Applications for Advanced Text Mining*. IntechOpen. https://doi.org/10.5772/51066

Castillo Molina, C. A., Gutierrez, R. E., & Solarte, O. (2015). Prototipo para el reconocimiento de entidades nombradas en el idioma Español. *2015 10th Computing Colombian Conference (10CCC)*, 364-371. https://doi.org/10.1109/ColumbianCC.2015.7333447

Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., & Xu, H. (2015). A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, *58*, 11-18.

Corpus lingüístico. (2022). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Corpus_ling%C3%BC%C3%ADstico&oldid=147071464

CORRALES, M., ANDRÉS, S., IZA, I., LIBELIA, J., UYAGUARI, I. U., & DANILO, A. (s. f.). *DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN CARRERA DE SOFTWARE*.

Costas, R., Moreno, L., & Bordons, M. (2008). Solapamiento y singularidad de MEDLINE, WoS e IME para el análisis de la actividad científica de una región en Ciencias de la Salud. *Revista espanola de documentación científica*, *31*(3), 327-343.

Descripción general de la API de Cloud Healthcare | API de Cloud Healthcare | Google Cloud. (s. f.). Recuperado 7 de febrero de 2023, de https://cloud.google.com/healthcare-api/docs/introduction?hl=es-419

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., & Krallinger, M. (2019). PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, 1-10. https://doi.org/10.18653/v1/D19-5701

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., & Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, *33*(14), i37-i48. https://doi.org/10.1093/bioinformatics/btx228

Ju, Z., Wang, J., & Zhu, F. (2011). Named Entity Recognition from Biomedical Text Using SVM. *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, 1-4. https://doi.org/10.1109/icbbe.2011.5779984

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139-172. https://doi.org/10.1080/23270012.2020.1756939

Kharwar, R. (2021, enero 6). "Medical Notes—An Underutilized Resource." *Google Cloud - Community*. https://medium.com/google-cloud/medical-notes-an-underutilized-resource-73c9e4b1a140

Leiva, J. (2022, junio 11). Ciencia de datos: Qué es y por qué es tan importante. *Ebac*. https://ebac.mx/blog/que-es-la-ciencia-de-datos

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, julio 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. ArXiv.Org. https://doi.org/10.48550/arXiv.1907.11692

Mahalakshmi, M., & Sundararajan, M. (2013). *Traditional SDLC Vs Scrum Methodology—A Comparative Study*. https://www.semanticscholar.org/paper/Traditional-SDLC-Vs-Scrum-Methodology-A-Comparative-Mahalakshmi-Sundararajan/7740829e70c028a75780d3b7bd034345beb940c4

Mariño, S. I., & Alfonzo, P. L. (2014). Implementación de SCRUM en el diseño del proyecto del trabajo final de aplicación. *Scientia et Technica, 2014, Año XIX, Vol. 19, No. 4, p. 413-418.* http://repositorio.unne.edu.ar/xmlui/handle/123456789/32471

Min, S. W., Wu, K., Huang, S., Hidayetoğlu, M., Xiong, J., Ebrahimi, E., Chen, D., & Hwu, W. (2021). *PyTorch-Direct: Enabling GPU Centric Data Access for Very Large Graph Neural Network Training with Irregular Accesses* (arXiv:2101.07956). arXiv. https://doi.org/10.48550/arXiv.2101.07956

O'Neill, C. (2022, abril 26). *An introduction to transformers and Hugging Face*. Medium. https://towardsdatascience.com/an-introduction-to-transformers-and-hugging-face-

13052ec9d72d

Peng, J., Zhao, M., Havrilla, J., Liu, C., Weng, C., Guthrie, W., Schultz, R., Wang, K., & Zhou, Y. (2020). Natural language processing (NLP) tools in extracting biomedical concepts from research articles: A case study on autism spectrum disorder. *BMC Medical Informatics and Decision Making*, *20*(11), 322. https://doi.org/10.1186/s12911-020-01352-2

PlanTL-GOB-ES/bsc-bio-ehr-es · Hugging Face. (s. f.). Recuperado 6 de febrero de 2023, de https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es

PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer · Hugging Face. (2022, noviembre 18). https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer

Quevedo-Marcos, B. (Borja). (2020). *Análisis de las herrramientas de procesamiento de lenguaje natural para estructurar textos médicos.* https://dadun.unav.edu/handle/10171/60003

Schwaber, K. (1997). SCRUM Development Process. En J. Sutherland, C. Casanave, J. Miller, P. Patel, & G. Hollowell (Eds.), *Business Object Design and Implementation* (pp. 117-134). Springer. https://doi.org/10.1007/978-1-4471-0947-1_11

Srivastava, A., Bhardwaj, S., & Saraswat, S. (2017). SCRUM model for agile methodology. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 864-869. https://doi.org/10.1109/CCAA.2017.8229928

Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H. (2012). Clinical entity recognition using structural support vector machines with rich features. *Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics*, 13-20. https://doi.org/10.1145/2390068.2390073

Vidal-Silva, C. L., Sánchez-Ortiz, A., Serrano, J., Rubio, J. M., Vidal-Silva, C. L., Sánchez-Ortiz, A., Serrano, J., & Rubio, J. M. (2021). Experiencia académica en desarrollo rápido de sistemas de información web con Python y Django. *Formación universitaria*, *14*(5), 85-94. https://doi.org/10.4067/S0718-50062021000500085

Villamizar Suaza, K., Tabares García, J. J., & Zapata Jaramillo, C. M. (2015). Mejora de historias de usuario y casos de prueba de metodologías ágiles con base en TDD. *Cuaderno Activa*, *7*, Art. 7.

Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical named entity recognition using deep learning models. *AMIA annual symposium proceedings*, *2017*, 1812.

Anexos