Modelo basado en Minería de textos y Procesamiento del Lenguaje Natural (NLP) para la gestión de artículos científicos – caso de estudio Nanopartículas

Cerón Ñauñay, María Belén

Departamento de Ciencias de la Computación

Carrera de Software

Trabajo de integración curricular, previo a la obtención del título de Ingeniera en Software

Ing. Gualotuña Álvarez, Tatiana Marisol, PhD.

06 de julio de 2022

CERON_TESIS_vFinal.docx

Scanned on: 18:48 July 5, 2022 UTC

Firmado electrónicamente por:
TATIANA MARISOL
GUALOTUNA
ALVAREZ

6.9%	40	16111
Overall Similarity Score	Results Found	Total Words in Text

Identical Words	398
Words with Minor Changes	434
Paraphrased Words	192
Omitted Words	1243

Departamento de Ciencias de la Computación

Carrera de Software

Certificación

Certifico que el trabajo de integración curricular: "**Modelo basado en Minería de textos y Procesamiento del Lenguaje Natural (NLP) para la gestión de artículos científicos – caso de estudio Nanopartículas**" fue realizado por la señorita **Cerón Ñauñay, María Belén,** 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 analizada 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.

Sangolquí, 06 de julio de 2022

Firma:

Firmado electrónicamente por:
TATIANA MARISOL
GUALOTUNA
ALVAREZ

.....

Ing. Gualotuña Álvarez, Tatiana Marisol, PhD.

C. C.: 1711498418

Departamento de Ciencias de la Computación

Carrera de Software

Responsabilidad de Autoría

Yo, **Cerón Ñauñay, María Belén**, con cédula de ciudadanía nº 1724266091, declaro que el contenido, ideas y criterios del trabajo de integración curricular: **Modelo basado en Minería de textos y Procesamiento del Lenguaje Natural (NLP) para la gestión de artículos científicos – caso de estudio Nanopartículas** es de mi 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.

Sangolquí, 06 de julio de 2022

Firma

.....

Cerón Ñauñay, María Belén

C. C.: 1724266091

Departamento de Ciencias de la Computación

Carrera de Software

Autorización de Publicación

Yo, **Cerón Ñauñay María Belén**, con cédula de ciudadanía no: 1724266091, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **Modelo basado en Minería de textos y Procesamiento del Lenguaje Natural (NLP) para la gestión de artículos científicos – caso de estudio Nanopartículas** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 06 de julio de 2022

Firma

.....

Cerón Ñauñay, María Belén

C. C.: 1724266091

Dedicatoria

A Dios por darme la fortaleza necesaria para seguir adelante cada día, por darme salud y vida para seguir cumpliendo mis sueños y alcanzar una de mis metas más anheladas como ahora la culminación de esta etapa de mi carrera.

A mis padres, por ser el pilar en mi vida, por ayudarme durante todo este tiempo y por el esfuerzo que han hecho para darme una educación de calidad, gracias a su amor, entrega y paciencia hoy empiezo una nueva etapa en mi vida.

A mis hermanos, por siempre apoyarme en todo momento, por alegrarme la vida y darme consejos cada día.

A todas las personas que formaron parte de este proceso y siempre me apoyaron.

Agradecimiento

Quiero agradecer primero a Dios, porque siempre me ha guiado por el camino del bien y porque me bendice y cuida cada día, me dio la fortaleza de seguir adelante cuando más lo necesitaba.

Les agradezco a mis padres que desde pequeña me enseñaron los valores y responsabilidades que hoy me han servido para ser la persona que soy. A mi mami por siempre brindarme su apoyo y por sus consejos tan valiosos, a mi papi porque cada día me enseña que el que persevera alcanza y por el amor que ambos me brindan. A mis hermanos y toda mi familia, gracias por todo.

Les agradezco a los docentes del departamento de Ciencias de la Computación de esta maravillosa universidad por haberme compartido sus enseñanzas y experiencias a lo largo de esta etapa académica. Además, quisiera agradecer de manera especial a la Ing. Tatiana Gualotuña por haber aceptado este reto conmigo y siempre apoyarme durante esta etapa y a lo largo de mi formación académica. Además, también le agradezco al Ing. Diego Marcillo por haberme ayudado y guiado para desarrollar este tema.

Quisiera agradecer a los amigos que me dio esta hermosa universidad, a Elian y Alex, gracias por las experiencias y grandes momentos que vivimos durante nuestra etapa de formación académica, por su ayuda y apoyo en todo momento. Finalmente, quisiera agradecer a mis más grandes amigas que me dio la vida, a Domi, Adri, Liss, Pao, gracias por apoyarme incondicionalmente.

Índice de Contenido

Índice de Tablas

Índice de Figuras

Resumen

La era digital ha marcado un antes y un después en la comunicación global, siendo una de las más relevantes la facilidad de compartir información, cada día se publican artículos científicos que son aportes muy importantes para la comunidad y estar al tanto de dichos avances es esencial, sin embargo, al existir tantos documentos digitales, el obtener información puntual suele ser una tarea que demanda mucho esfuerzo y tiempo. Para el caso de estudio se tomó en consideración los problemas que presentan los investigadores del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE, donde a través de un análisis exploratorio, se determinó que existe una inadecuada gestión manual de la información en artículos científicos sobre nanopartículas. Por esta razón, el proyecto tiene como propósito desarrollar un modelo machine learning para la gestión de artículos científicos a través del cual se puedan establecer patrones de información que aporten al conocimiento científico.

Este proyecto se elaboró basándose en la metodología Design Science Research (DSR) y por medio de este, se definió e implementó el modelo machine learning para la generación de patrones de comportamiento en las investigaciones relacionadas con nanopartículas. Finalmente, se realizó la validación del modelo utilizando métricas de evaluación de rendimiento de clasificadores en tareas de extracción de datos y, además, al poner en marcha el modelo se obtuvieron resultados positivos donde se contrastaron los datos arrojados por el modelo y las anotaciones extraídas manualmente, determinándose que el modelo puede ser utilizado como una herramienta de apoyo para las investigaciones en el campo de las nanopartículas.

Palabras clave: machine learning, procesamiento del lenguaje natural, minería de textos, extracción automática de datos, nanopartículas.

Abstract

The digital era has marked a before and after in global communication, being one of the most relevant the ease of sharing information, every day scientific articles are published that are very important contributions to the community and being aware of these advances is essential, however, with so many digital documents, obtaining timely information is usually a task that demands a lot of effort and time. For the case study, we took into consideration the problems presented by researchers in the area of Biotechnology at the Universidad de las Fuerzas Armadas – ESPE, where through an exploratory analysis, it was determined that there is an inadequate manual management of information in scientific articles on nanoparticles. For this reason, the project aims to develop a machine learning model for the management of scientific articles through which information patterns that contribute to scientific knowledge can be established.

This project was developed based on the Design Science Research (DSR) methodology and through this, the machine learning model was defined and implemented for the generation of behavioral patterns in research related to nanoparticles. Finally, the validation of the model was performed using classifier performance evaluation metrics in data extraction tasks and, in addition, when the model was implemented, positive results were obtained where the data yielded by the model and the manually extracted annotations were contrasted, determining that the model can be used as a support tool for research in the field of nanoparticles.

Keywords: machine learning, natural language processing, text mining, automatic data mining, nanoparticles.

Capítulo I

Introducción

Antecedentes

El ser humano siempre ha buscado métodos que faciliten su forma de comunicarse; a lo largo de los años estos métodos han ido evolucionando y con el crecimiento acelerado de las tecnologías, como el Internet, ha permitido compartir información desde cualquier parte del mundo.

Actualmente, el mundo se rige por los avances tecnológicos. Con el surgimiento de la era digital, la información se convirtió en el eje principal de esta nueva era; los medios informativos han actualizado procesos como el almacenamiento y distribución de la información, ya sea en formato de texto, imagen, sonido, entre otras (Jódar, 2010).

Hoy en día, mucha información se encuentra disponible en bases de datos; precisamente, existen bases de datos académicas que proporcionan funciones de acceso a la información, documentos publicados o bien por editores científicos-académicos o por asociaciones científicas o universidades (Codina, Morales, Rodríguez, & Pérez, 2020).

Los documentos publicados son denominados artículos científicos. Este tipo de documentos contribuye a la construcción del conocimiento, ya que a través de estos los investigadores comparten los resultados que han obtenido en sus investigaciones. De esta manera, dichos documentos son publicados en bases de datos como, por ejemplo, Scopus, IEEE Xplore, Web of Science, entre otras (Codina, 2020). Estas bases son capaces de almacenar mucha información, sin embargo, esta información crece exponencialmente y cada día resulta más difícil encontrar información específica y estar actualizado de los temas de interés (Morales & Tobar, 2020).

En todo el mundo, al año se publican alrededor de tres millones de artículos científicos, y aunque parece una cifra positiva no todos los artículos científicos exponen información de calidad porque algunos presentan malas prácticas y resultados incorrectos (Lorite, 2019). Por esta razón muchos investigadores se quejan de la sobreabundancia de información y por consiguiente provoca algunos problemas dentro de los procesos de investigación (Russell, 2001).

Al realizar una investigación, uno de los procesos más importantes es la revisión sistemática de literatura dado que facilita la recopilación, síntesis y análisis de artículos científicos. No obstante, este proceso es largo, laborioso, requiere dedicación y tiempo puesto que se debe revisar y validar la calidad de estos documentos manualmente (Fredes, 2017).

En la Universidad de las Fuerzas Armadas – ESPE, al año cada departamento realiza trabajos de investigación y publicaciones de artículos científicos. Según los datos recolectados por parte de la universidad, entre los años 2016 y 2020, la media de publicaciones dentro de bases de datos mundiales es de 335 y dentro de bases de datos regionales es de 87 artículos científicos (Universidad de las Fuerzas Armadas ESPE, 2020).

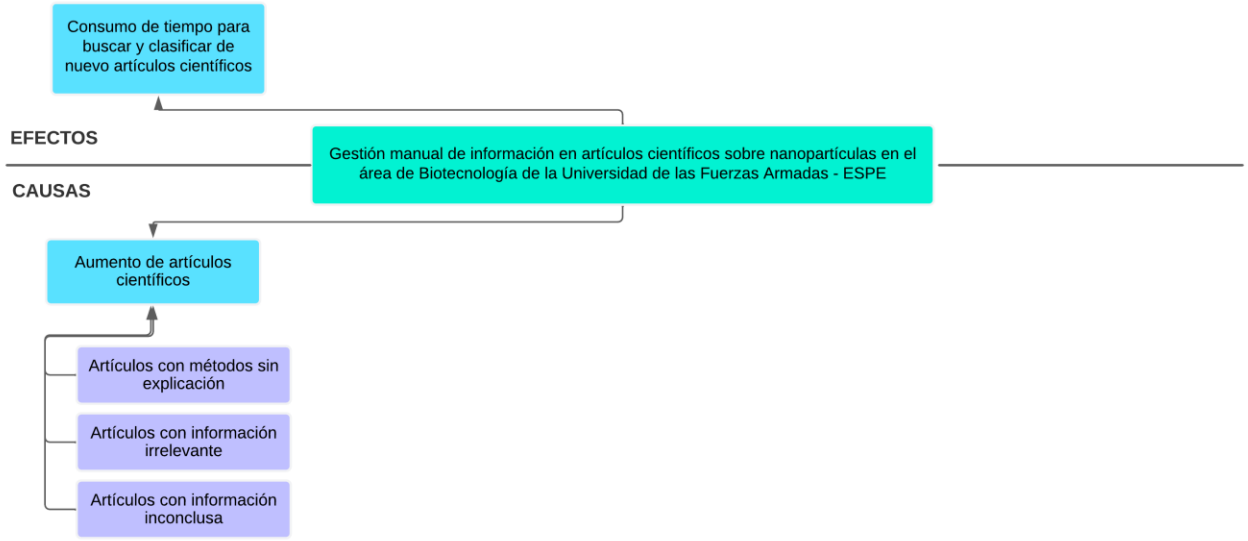
En particular, el área de Biotecnología también es parte de esta media de publicaciones, sin embargo, los investigadores han encontrado problemas parecidos a los descritos previamente, especialmente en la etapa de recolección de información. Entonces, la Universidad de las Fuerzas Armadas – ESPE al ser una universidad reconocida a nivel nacional y regional, dado el contexto descrito con anterioridad, se vio factible realizar el estudio en el área de Biotecnología de la universidad.

Problemática

En esta investigación se abordó el problema de la gestión manual de información en artículos científicos sobre nanopartículas en el área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE. Para la identificación de la problemática se aplicó entrevistas[1] al personal del laboratorio de caracterización de nanomateriales del Departamento de Ciencias de la Vida de la universidad.

Figura 1

Aumento de artículos científicos en bases de datos académicas



Nota. Este gráfico representa la primera causa de la problemática.

Los resultados de la investigación demuestran que la primera causa del problema identificado es el **aumento de artículos científicos** en las bases de datos académicas (ver figura 1). Como se mencionó anteriormente, en la actualidad se publican muchos artículos científicos y la existencia de tantos no garantiza que todos sean artículos de calidad. Esto es comprobado por los entrevistados donde mencionan que, al leer los artículos científicos,

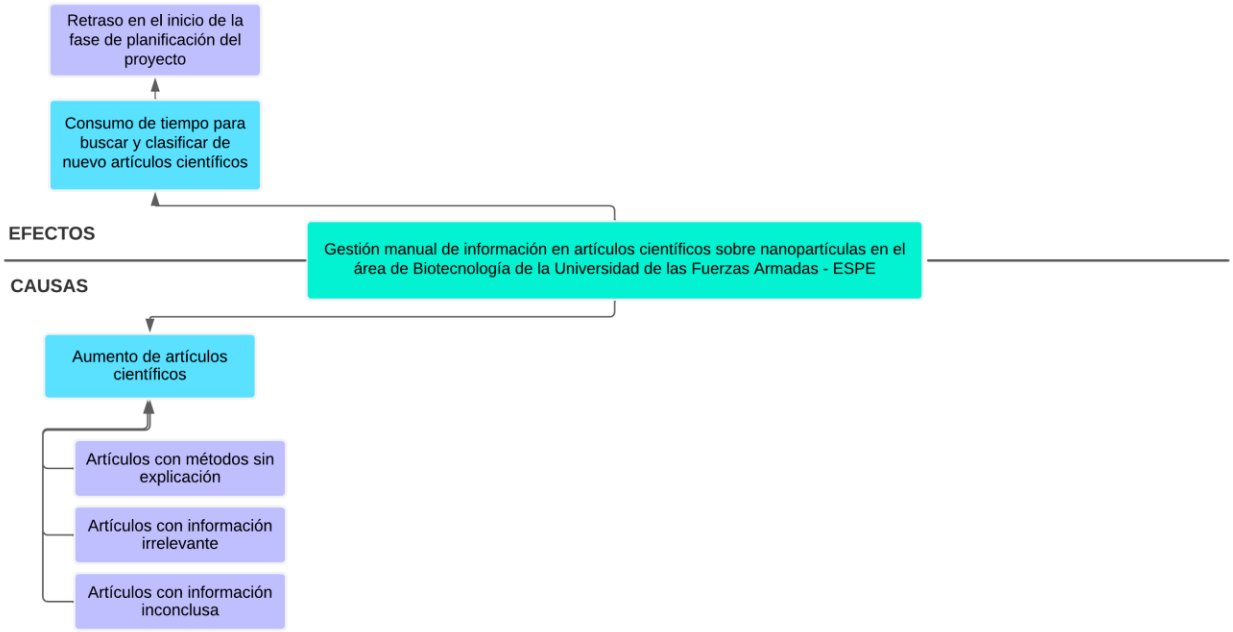
[1] La entrevista se encuentra disponible en el Apéndice 1

algunos contienen **información irrelevante, con métodos sin explicación suficiente y con información inconclusa**.

Cuando los investigadores se encuentran con este tipo de artículos científicos, deben regresar a las fases iniciales, en particular a la fase de recolección de información para encontrar artículos que complementen a los artículos previos o en su defecto buscar nuevos. Todo este proceso genera que se **consuma más tiempo para clasificar nuevos artículos científicos** y por consiguiente **retrasa la fase de planificación del proyecto** (ver figura 2).

Figura 2

Efectos del aumento de artículos científicos



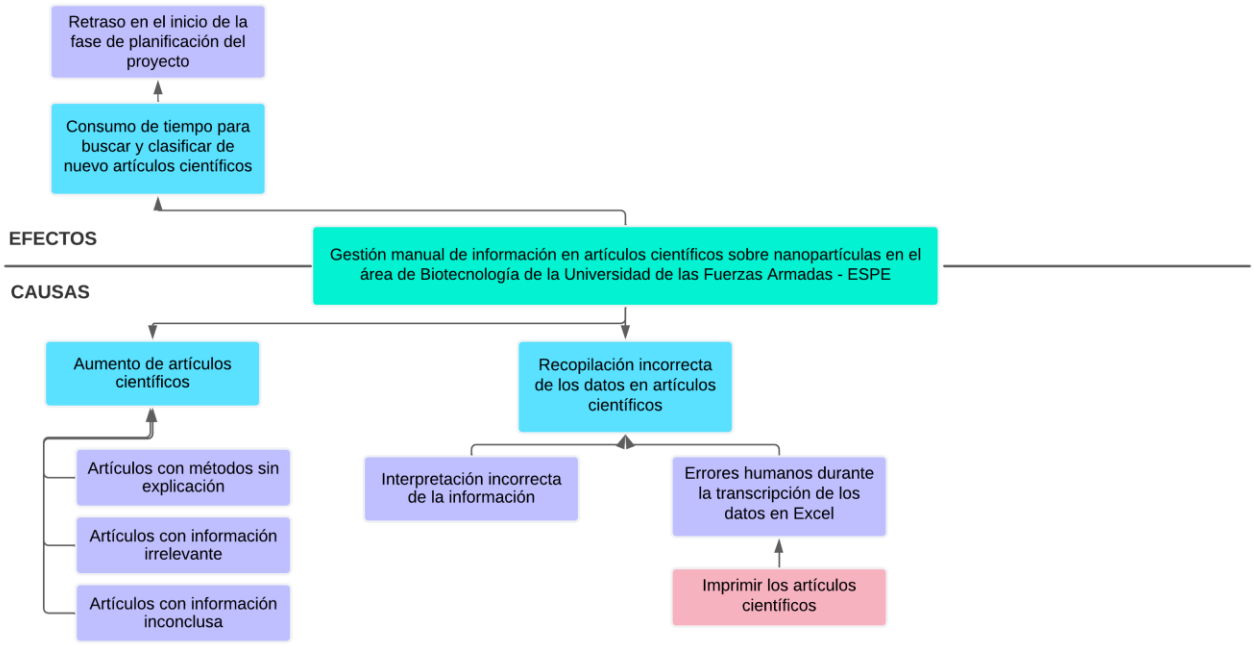
Nota. Este gráfico representa los primeros efectos causados por la problemática en cuestión.

La fase de recopilación de información conlleva dos etapas principales. La primera es el análisis del estado del arte, donde los investigadores buscan los artículos más representativos dentro del campo de la nanotecnología. La segunda etapa tiene que ver con la recolección de datos relevantes encontrados en los artículos científicos. Sin embargo, esta fase es susceptible

a errores humanos. Uno de estos es la **interpretación incorrecta de la información** ya que cada persona interpreta de manera diferente los datos. Por otra parte, los entrevistados mencionaron que se suele imprimir los artículos científicos para facilitar la lectura, y posteriormente se ingresan los datos encontrados a una hoja de Excel. El problema de ingresar manualmente los datos a Excel es que existen **errores al transcribir dichos datos** ya que las personan pueden tener errores de tipeo (ver figura 3).

Figura 3

Recopilación incorrecta de los datos en artículos científicos



Nota. Este gráfico representa la segunda casusa de la problemática en cuestión.

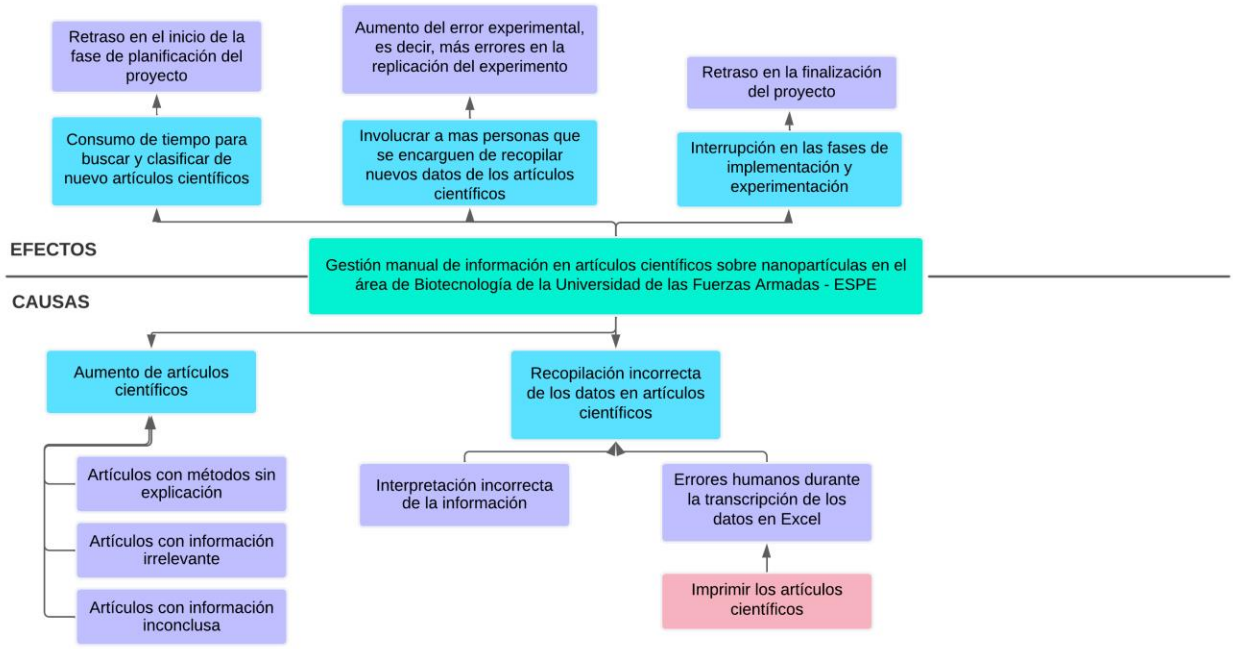
Adicionalmente, se identificó que un investigador tarda alrededor de 2 a 4 semanas en extraer datos de los artículos científicos que han sido seleccionados, si esta tarea lo realizan dos personas tarda la mitad del tiempo y así sucesivamente. Sin embargo, dentro de un proyecto de investigación sobre nanopartículas se trata de involucrar a la menor cantidad de personas posibles porque mientras más personas, más errores existirán en la replicación de los experimentos. Entonces, cuando existen errores durante la extracción de los datos, por el

tiempo que se estableció para finalizar el proyecto se tiene que **involucrar a más personas para que se encarguen de recopilar o corregir nuevos datos de los artículos científicos** lo que **aumenta el error experimental** (ver figura 4).

De forma similar este problema ocasiona que se **retrasen las fases de implementación y experimentación** ya que, sin datos precisos como métodos, cantidades de las sustancias, entre otras, no es posible replicar los experimentos y por consiguiente existirán **retrasos para finalizar el proyecto**.

Figura 4

Efectos del aumento de artículos científicos



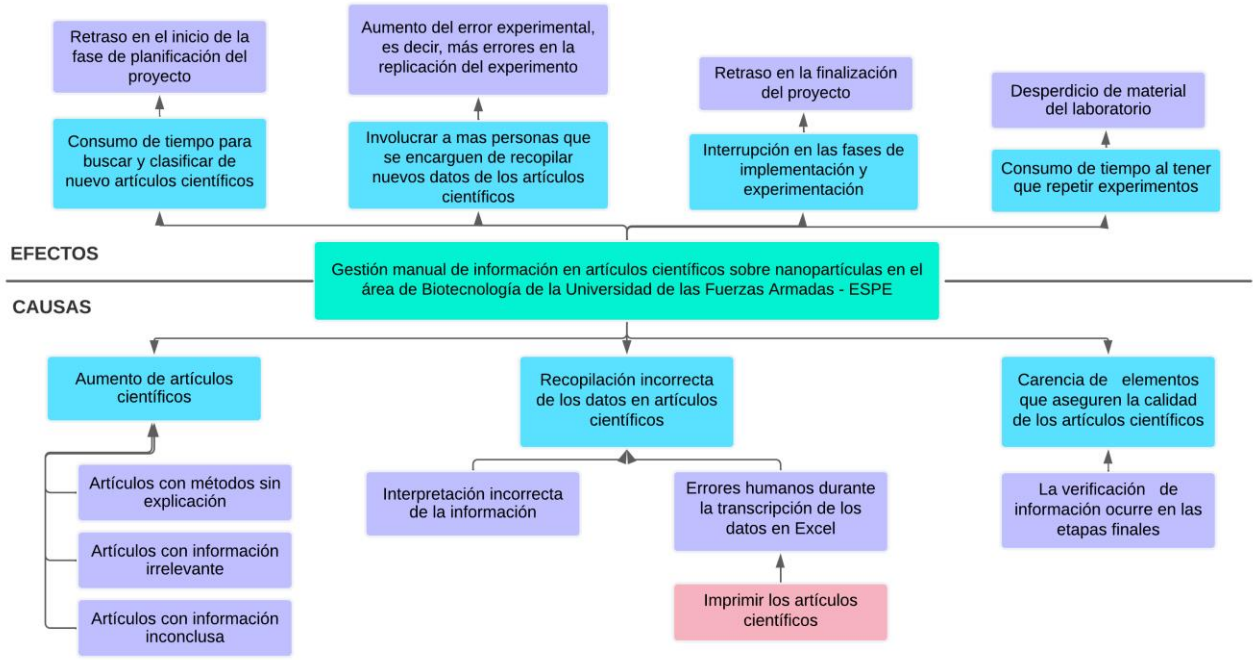
Nota. Este gráfico representa los segundos efectos causados por la problemática en cuestión.

Finalmente, otro de los resultados que se encontró es que los investigadores **carecen de elementos que aseguren la calidad de los artículos científicos** dado que suelen verificar los datos extraídos a la par que realizan los experimentos ya que con su experiencia pueden reconocer si algún dato es incorrecto, sin embargo, cuando se tratan de datos numéricos como

cantidades, mediciones y demás, esto no es posible a simple vista. Para descubrir si algún dato es incorrecto se necesita analizar los resultados de los experimentos y con técnicas estadísticas se los validan; si se obtienen diferencias estadísticas muy grandes significa que existe un error en la extracción de los datos y por consiguiente se tendrá que **repetir el experimento** y esto ocasiona **desperdicio de tiempo y material de laboratorio** (ver figura 5)**.**

Figura 5

Árbol completo de causa y efecto de la problemática



Nota. Este gráfico sintetiza las causas y efectos de la problemática en cuestión.

Justificación

La recolección de datos es una tarea fundamental para procesos de investigación y experimentación, ya que ayuda a los investigadores a comprender mejor el contenido de un artículo científico. No obstante, con los resultados que se obtuvieron en el área de Biotecnología se ve necesario implementar un modelo capaz de extraer automáticamente datos

relevantes dentro de artículos científicos, particularmente utilizando un caso de estudio de Nanopartículas; debido a una amplia variedad de aplicaciones potenciales en diversos campos (Medina, Galván, & Reyes, 2015).

Objetivos

Objetivo General. Desarrollar un modelo machine learning para la gestión de artículos científicos a través del cual se pueda establecer patrones de información que aporten al conocimiento científico.

Objetivo Específico.

i. Realizar una revisión de literatura que permita determinar técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP), adecuadas para el procesamiento, etiquetado y extracción de datos.

ii. Definir e implementar el modelo machine learning para la generación de patrones de comportamiento en las investigaciones relacionadas con nanopartículas.

iii. Realizar la validación del modelo utilizando métricas de evaluación de rendimiento de clasificadores en tareas de extracción de datos.

Alcance

Este proyecto tiene como alcance implementar el modelo basado en técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP) para automatizar tareas de recolección de datos en artículos científicos de Nanopartículas. Además, realizar una comparativa entre los datos recolectados manualmente y los obtenidos por el modelo machine learning.

A continuación, se muestra la matriz de congruencia metodológica que se aplicó para determinar la revisión de literatura del proyecto. En esta se muestran los objetivos específicos con sus respectivas preguntas de investigación.

Tabla 1

Matriz de congruencia Metodológica

	Objetivo Específico	**Preguntas de Investigación**
i.	Realizar una revisión de literatura que permita determinar técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP), adecuadas para el procesamiento, etiquetado y extracción de datos.	**RQ1.** ¿Qué soluciones presentan los estudios consultados sobre la extracción de datos? **RQ2.** ¿Qué herramientas machine learning permiten encontrar patrones en los datos? **RQ3.** ¿Qué técnicas se utilizan para realizar el etiquetado de los datos?
ii.	Definir e implementar el modelo machine learning para la generación de patrones de comportamiento en las investigaciones relacionadas con nanopartículas.	**RQ4:** ¿Qué técnicas de extracción, transformación y limpieza son necesarias ejecutar, preparar los datos relacionados a textos no estructurados? **RQ5:** ¿Cuáles son los criterios de selección para definir el modelo implementado? **RQ6:** ¿Cuáles son los parámetros de ejecución utilizados y definidos para la implementación del modelo?
iii.	Realizar la validación del modelo utilizando métricas de evaluación de	**RQ7:** ¿Qué métricas se utilizarán para validar el modelo machine learning?

Objetivo Específico	Preguntas de Investigación
rendimiento de clasificadores en tareas de extracción de datos.	**RQ8:** Con los patrones encontrados ¿se puede apoyar a los investigadores? **RQ9:** ¿Qué niveles de asertividad y confianza ha generado el modelo?

Nota. Esta tabla muestra preguntas de investigación definidas para el proyecto.

Hipótesis

Un modelo machine learning basado en técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP) permite automatizar tareas de recolección de datos en artículos científicos de Nanopartículas.

Capítulo II

Estado del Arte

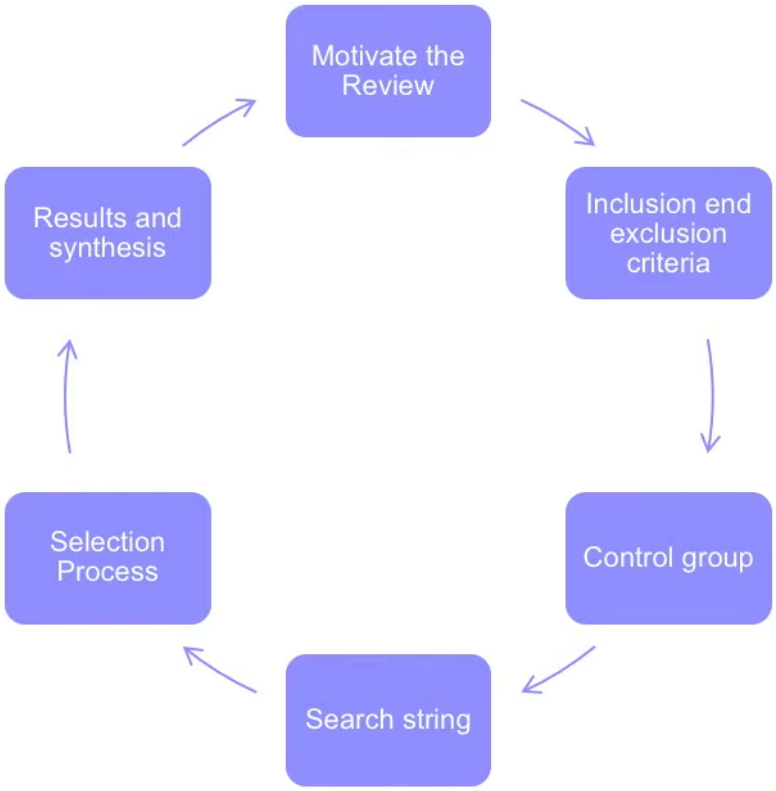
El estado del arte es una revisión de literatura donde, a través de una serie de pasos, se analizan trabajos preliminares que han sido planteados por otros autores sobre un tema en específico (Budgen, y otros, 2007).

Para este estudio, la técnica que se empleó fue la revisión preliminar de literatura (Fonseca, 2020), esto con el fin de recolectar trabajos que contribuyan a la validación de la propuesta del modelo machine learning para la gestión de artículos científicos. Además, para realizar esta tarea se utilizó la base de datos digital PubMed (PubMed, 2022) por su amplio catálogo de artículos científicos sobre biomedicina y ciencias de la vida, esto debido al contexto del estudio.

La revisión preliminar de literatura es un proceso iterativo que se compone de varias fases (ver figura 6). La primera fase es la motivación de la revisión, donde se identifica el problema. Una vez determinado el problema se definen los criterios de inclusión y exclusión, estos se utilizan para conocer las características que poseerán los estudios que se están buscando. Después, se conforma un grupo de control con los trabajos que cumplieron los criterios antes mencionados, de estos trabajos saldrán los términos más relevantes que posteriormente servirán para estructurar la cadena de búsqueda. Una vez validada la cadena de búsqueda, se realiza el proceso de selección de estudios primarios, que son los más representativos. Finalmente se reportan los resultados, es decir, se elabora el estado del arte (Fonseca, 2020).

Figura 6

Fases de una revisión preliminar de literatura



Nota. Este gráfico sintetiza el proceso para elaborar una revisión preliminar de literatura. Adaptado de *Preliminary Literature Review Theory - Video 1,* por R. Fonseca, 2020, Youtube (https://youtu.be/3zcY87cV0YQ). Derechos de autor 2020 por R. Fonseca.

Planteamiento de la revisión de literatura

Después de identificar el estado actual y la problemática que se detalló en el capítulo anterior, se continua con el objetivo referente a la revisión de literatura y posteriormente el planteamiento de las preguntas de investigación que serán contestadas en esta sección. Concretamente, corresponde al objetivo específico 1 y las preguntas de investigación 1, 2, 3.

Criterios de inclusión y exclusión

Los criterios de inclusión y exclusión son una de las fases más importantes ya que ayudan a delimitar la extensión que tendrán los artículos científicos que se buscan. Es

fundamental hacer explicito lo que se está buscando porque de lo contrario se obtendrán cantidades de artículos científicos inmanejables.

Los criterios de inclusión (CI) corresponden a las características que se busca dentro de un artículo científico, por otra parte, los criterios de exclusión (CE) se refieren a las características que conducirán al rechazo de un artículo científico. A continuación, se presentan los criterios de inclusión definidos para este estudio:

- *CI1:* Estudios que presenten soluciones con técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP) para la extracción de datos en textos no estructurados.
- *CI2:* Estudios que especifiquen herramientas machine learning que permitan encontrar patrones dentro de los datos.
- *CI3:* Estudios que detallen los métodos que utilizaron para etiquetar los datos.
- *CI4:* Estudios que indiquen como se realiza el preprocesamiento de los datos, especialmente sobre la limpieza de los datos.

Los artículos que serán excluidos serán aquellos:

- *CE1:* Estudios que no presenten soluciones en textos no estructurados para la extracción de datos.
- *CE2:* Estudios publicados antes del año 2005.
- *CE3:* Estudios que utilicen software que necesite algún tipo de licencia para su uso.

Grupo de Control

El grupo de control (CG) está conformado por estudios que cumplen completamente con los criterios de inclusión y exclusión previamente definidos. Estos estudios son propuestos

por los investigadores involucrados y es recomendable realizar este proceso entre dos o más personas para discutir los estudios más adecuados que conformarán el grupo de control.

Tabla 2

Grupo de control

Código	Título	Citas	Términos relevantes
CGS1[2]	SciNER: Extracting Named Entities from Scientific Literature	(Hong, Tchoua, Chard, & Foster, 2020)	Scientific Literature, model, text extraction, bioinformatic, biomedical, architecture, science
CGS2	ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature	(Swain & Cole, 2016)	Scientific articles, Scientific Literature, text extraction, information management, toolkit, model, chemical information, bioinformatic
CGS3	@Note: A workbench for Biomedical Text Mining	(Lourenço, y otros, 2009)	Information retrieval, Information extraction, literature curation, biomedical, model, bioinformatic

Nota. Esta tabla muestra los estudios que conforman el grupo de control.

[2] CGS: Estudio del Grupo de Control

Una vez conformado el grupo de control, se analiza cada estudio para así obtener los términos más representativos. Estos términos son fundamentales ya que sirve como pauta inicial para la posterior construcción de la cadena de búsqueda.

Cadena de búsqueda

La cadena de búsqueda se conforma utilizando los términos relevantes que se encontraron en los estudios del grupo de control y sirve para obtener estudios relacionados a la temática en cuestión en las diferentes bases de datos.

Una vez identificado los términos relevantes se procede a conformar contextos, es decir, formar pequeños grupos de palabras relacionadas entre sí que serán agrupadas por conectores lógicos (“OR”, “AND”). De esta manera se conformaron 3 contextos que son los siguientes:

- Donde se realiza el estudio.
- Descripción de las tareas que se realizan en el estudio.
- Herramientas implementadas en el estudio.

Para la validación de la cadena de búsqueda se consideró los siguientes aspectos:

- El número de estudios es manejable.
- Títulos y resúmenes relacionados al tema.
- Se cumple con los criterios de inclusión y exclusión.
- Dentro de los resultados obtenidos con la cadena de búsqueda se encuentran la mayoría de los estudios del grupo de control.

Una vez conformada la cadena de búsqueda se procede a probarla, de esta manera se genera un proceso iterativo de prueba y error hasta encontrar la cadena que mejor se acople a

las características antes mencionadas. A continuación, se muestran las cadenas utilizadas durante este proceso y que fueron probadas en la base digital PubMed:

Tabla 3

Refinamiento de la cadena de búsqueda

Cadena de Búsqueda	Número de artículos
(("scientific articles" OR "scientific literature") AND ("text extraction" OR "information management" OR "information retrieval") AND ("framework" OR "architecture" OR "toolkit")**)**	6
(("biotechnology" OR "nanoparticles") AND ("scientific articles" OR "Scientific Literature") AND ("text extraction" OR "information management" or "information retrieval") AND ("framework" OR "model" OR "architecture" OR "toolkit")**)**	1
(("scientific articles" OR "Scientific Literature" OR "bioinformatic") AND ("text extraction" OR "information management" or "information retrieval") AND ("framework" OR "model" OR "architecture" OR "toolkit")**)**	16
(("scientific articles" OR "Scientific Literature" OR "bioinformatic") AND ("information extraction" OR "text extraction" OR "information management" or "information retrieval") AND ("model" OR "architecture" OR "toolkit")**)**	31

Nota. Esta tabla muestra la evolución que tuvo la cadena de búsqueda ideal.

En la tabla anterior se puede observar las fases que tuvo el proceso de prueba, en este caso se seleccionó la cadena de búsqueda que arrojó 31 estudios, esto debido a que contiene más artículos relacionados a la temática.

Proceso de selección

Para este proceso se toman en cuenta 3 tipos de estudios; Estudios Candidatos (EC), Estudios Relevantes (ER) y Estudios Primarios (EP).

Los EC se los obtienen a través de la cadena de búsqueda, en este caso se tienen 31 EC. Posteriormente se realiza un proceso de filtrado, es decir, se lleva a cabo una lectura rápida de los EC tomando en cuenta el título, resumen y palabras clave. Una vez realizado este proceso se obtienen los ER, en este caso se encontraron 4 ER, es decir, no existe mucha investigación relacionada al tema.

Después, para seleccionar los estudios que conformaran los EP se debe descargar los ER y leer cada uno de ellos. Finalmente, los ER se convierten en EP y junto con los estudios del grupo de control, se obtuvieron un total de 7 EP.

Tabla 4

Estudios Primarios

Código	Título	Citas
EP1	SciNER: Extracting Named Entities from Scientific Literature	(Hong, Tchoua, Chard, & Foster, 2020)
EP2	ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature	(Swain & Cole, 2016)
EP3	@Note: A workbench for Biomedical Text Mining	(Lourenço, y otros, 2009)

Código	Título	Citas
EP4	A pre-training and self-training approach for biomedical named entity recognition	(Gao, Kotevska, Sorokine, & Christian, 2021)
EP5	Bio-semantic relation extraction with attention-based external knowledge reinforcement	(Li, Lian, Ma, Zhang, & Li, 2020)
EP6	Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets	(Vashishth, Newman-Griffis, Joshi, Dutt, & Rosé, 2021)
EP7	Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks	(Wei, Chen, Xu, He, & Gui, 2016)

Nota. Esta tabla muestra los estudios primarios seleccionados.

Resumen de los Estudios Primarios

EP1: SciNER: Extracting Named Entities from Scientific Literature. Este artículo trata el tema del desarrollo de una herramienta automatizada que facilite la obtención de información relevante en artículos científicos. Los autores presentan un modelo basado en redes neuronales, específicamente en redes LSTM bidireccionales denominado “SciNER”. Este es un modelo generalizado el cual es capaz de adaptarse a cualquier temática dentro de artículos científicos. Este modelo combina varias técnicas como “word embedding”, “subword embeddings” y, además, esta alimentado por una fuente externa de conocimiento como “DBpedia”. Después, realizan un experimento en dos áreas como lo es las ciencias naturales para obtener nombres de polímeros y en las ciencias sociales para obtener nombres de localidades y organizaciones. Finalmente, primero comparan su modelo con

ChemDataExtractor donde los resultados obtenidos muestran una precisión del 0.909 para la identificación de nombres de polímeros y después para la identificación de nombres de localidades y organizaciones comparan su modelo con un clasificador KNN obteniendo una puntuación del 0.847 en precisión.

EP2: ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. En este artículo se detallan todas las fases necesarias para desarrollar un conjunto de herramientas para la extracción automática de información química dentro de artículos científicos y que además se encuentra disponible en su página web oficial. Los autores presentan un “toolkit” denominado “ChemDataExtractor” que es capaz de extraer entidades químicas y sus respectivas propiedades. Este sistema combina varias técnicas de Procesamiento del Lenguaje Natural como “tokenization”, “part-of-speech tagging”, “name entity recognition” y “phrase parsing”. Una característica interesante de su sistema es la capacidad de procesar tablas en artículos científicos y demás información. Finalmente, los autores realizan un experimento con algunos artículos científicos obteniendo una media de precisión de extracción del 90% en todas las entidades que utilizaron.

EP3: @Note: A workbench for Biomedical Text Mining. Este articulo presenta una plataforma para facilitar tareas de investigación en artículos científicos el cual puede ser utilizada por 3 tipos de usuarios: Biólogos, Mineros de textos y Desarrolladores de software. Esta plataforma se denomina “@Note” y consta de 4 módulos principales los cuales son: (1) Módulo de Recuperación de Documentos, encargado de buscar información dentro de la base digital PubMed; (2) Módulo de Conversión y Estructuración de Documentos, encargado de convertir los archivos PDF en archivos de texto, además de implementar herramientas para la limpieza de los datos; (3) Módulo de Procesamiento del Lenguaje Natural, encargado del preparamiento de los datos para la futura construcción de modelos NER para la extracción automática de datos dentro de estos textos y otras tareas relacionadas al NLP; (4) Módulo de

Minería de texto, encargado de la construcción y evaluación de modelos NER. Los autores mencionan que esta herramienta fue diseñada para que los tipos de usuarios antes mencionados puedan utilizar esta plataforma fácilmente, además para los desarrollos de software se les permite agregar otros módulos si así lo desearan.

EP4: A pre-training and self-training approach for biomedical named entity recognition. Este artículo trata sobre el desarrollo de una herramienta automatiza para extraer información biomédica relevante como por ejemplo la detección de nombres de proteínas, genética, drogas, químicos, enfermedades y especies utilizando un corpus conformado por varios datos. Además, los autores evalúan la efectividad de combinar el aprendizaje por transferencia con el aprendizaje semi-supervisado para entrenar un modelo NER basado en redes BiLSTM-CRF y BERT esto con el objetivo de tener la capacidad de entrenar dichos modelos sin tantos datos etiquetados. En esta parte se concluyó que no hay mucha precisión en los modelos que son entrenados con datos sin etiquetar, pero aquí viene el problema de que para obtener modelos eficientes se necesita de corpus de datos etiquetados por profesionales y obtener este tipo de datos es bastante complejo de conseguir.

EP5: Bio-semantic relation extraction with attention-based external knowledge reinforcement. Este articulo detalla el proceso de extracción de información relevante en documentos científicos sobre biología. Los autores presentan un modelo basado en redes neuronales profundas BiLSTM aprovechado algunas bases del conocimiento como UnitProt y BioModels que se utilizaron para alimentar de datos al modelo. Este modelo es evaluado utilizando el corpus BioNLP y BioCreative que son un conjunto de datos con anotaciones manuales. En particular, los autores se enfocaron en la detección de entidades y relaciones semánticas de proteínas y genes. Finalmente, luego de realizar las respectivas pruebas de rendimiento al modelo, los autores concluyen que, si bien usar este tipo de enfoque es

generalizado y podría utilizarse con otro tipo de datos, mencionan que este modelo debe seguir siendo mejorado.

EP6: Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. Este articulo presenta un módulo para la predicción semántica con la posibilidad de utilizarlo en tareas de extracción de texto biomédico. Este módulo se denomina “MEDTYPE” y utiliza los beneficios de BERT (Representaciones de Codificador Bidireccional de Transformadores), un algoritmo desarrollado por Google. Además, crearon dos corpus de datos denominados “WikiMed” y “PubMedDS” con los cuales entrenaron su módulo. En la parte de experimentación los autores realizaron pruebas de extracción de información biomédica a partir de cuatro corpus diferentes donde se identificaron diferentes entidades (anatomía, proteínas, genética, enfermedades, entre otros) donde utilizaron cinco kits de herramientas de NLP biomédica junto con su módulo de predicción de tipo semántica donde concluyeron que haciendo uso de este módulo las herramientas de extracción de información mejoran notablemente su precisión.

EP7: Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. Este articulo presenta el desarrollo de un modelo para la detección de nombres de enfermedades en artículos científicos. Este modelo general parte de dos modelos NER por separado. El primero se basa en un modelo de campos aleatorios condicionales y el segundo se basa en redes neuronales recurrentes bidireccionales. Una vez que las entidades son reconocidas por cada uno de los modelos estos a su vez se unen a otro, que es un clasificador encargado de unir ambas respuestas. El propósito de realizar dos modelos y uno general es para tener más precisión en la extracción de nombres de enfermedades y esto es validado con las pruebas realizadas obteniendo una precisión del 84.28%.

Respuestas a las preguntas de investigación

¿Qué soluciones presentan los estudios consultados sobre la extracción de datos? Las soluciones presentadas por todos los artículos antes mencionados para la extracción de información dentro de fuentes no estructuradas como artículos científicos es la utilización de NER (Named Entity Recognition). Además, NER es parte del campo del Procesamiento del Lenguaje Natural (NLP) y minería de textos.

¿Qué herramientas machine learning permiten encontrar patrones en los datos? Para encontrar patrones dentro de los datos cada artículo científico antes mencionado brinda algunas soluciones como por ejemplo el uso de modelos basados en BiLSTM, CRF, BERT o RNN.

Tabla 5

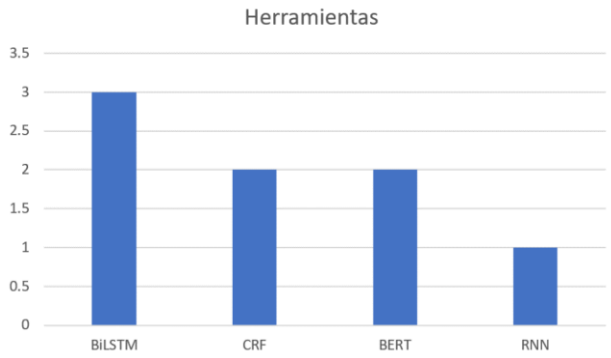
Soluciones presentadas por los Estudios Primarios

Modelo	Estudios
BiLSTM	Tres investigaciones: EP1, EP4, EP5.
CRF	Dos investigaciones: EP4, EP7.
BERT	Dos investigaciones: EP4, EP6.
RNN	Una investigación: EP7.

Nota. Esta tabla muestra las soluciones presentadas por cada EP.

Figura 7

Herramientas utilizadas por los estudios primarios

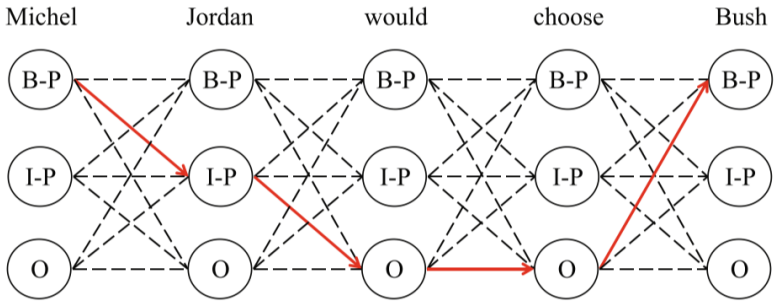


Nota. Este gráfico muestra las herramientas utilizadas en los EP.

¿Qué técnicas se utilizan para realizar el etiquetado de los datos? En general todos los artículos científicos mencionan que para etiquetar los datos primero se tiene que pasar por un proceso denominado preprocesamiento de datos y posteriormente se realiza el "BIO encoding" (Beginning-Inside-Outside) que es un método general para etiquetar entidades y de esta manera poder identificar dichas entidades dentro de corpus de textos.

Figura 8

Esquema BIO encoding



Nota. Este gráfico muestra como está estructurado el método BIO encoding para etiquetar textos. Adaptado de *Enhanced Sequence Labeling Based on Latent Variable Conditional Random Fields* (p. 434), por C. Lin, Y. Shao, J. Zhang y U. Yun, 2020, Neurocomputing.

Resumen general y conclusiones del estado del arte

El estado del arte permitió identificar soluciones, métodos y herramientas utilizadas actualmente para resolver problemas de extracción de información dentro de artículos científicos relacionados a biomedicina. Sin embargo, la mayoría de los artículos presentan soluciones para nombres de proteínas, genes, elementos químicos, entre otros, que son temas muy recurrentes y por lo tanto ya se cuenta con corpus de datos etiquetados lo cual facilita el entrenamiento de los modelos machine learning.

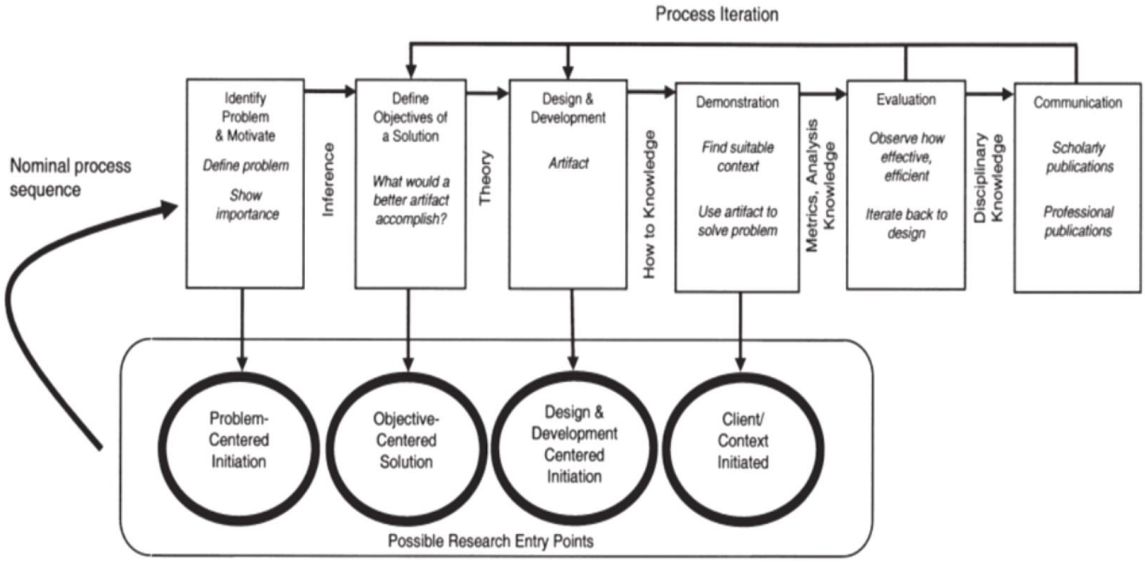
Asimismo, la mayoría de los artículos hacen comparaciones con otros modelos basados en los corpus ya etiquetados y no presentan en si una solución más específica como por ejemplo en el tema de nanopartículas. Por lo tanto, considerando estos antecedentes, se puede continuar con la investigación para el desarrollo de una herramienta machine learning que sea capaz de identificar más entidades relacionadas a nanopartículas.

Metodología

Para el desarrollo de esta investigación se consideró la aplicación de la metodología Design Science Research (DSR), esto debido al contexto del problema y por estar dentro del ámbito de la tecnología. Esta metodología es un paradigma de resolución de problemas que busca mejorar el conocimiento humano a través de la creación de artefactos innovadores. En pocas palabras, DSR busca mejorar las bases de conocimiento científico y tecnológico a través de la creación de artefactos innovadores que resuelven problemas y mejoran el entorno en el que se instancian (Brocke, Maedche, & Hevner, 2020). Una de las ventajas que presenta esta metodología es que es iterativa, es decir, si ocurre algún problema se puede regresar al paso anterior. A continuación, se detallan cada una de las fases y en la figura 9 se pueden observar las mismas.

Figura 9

Fases de la metodología DSR



Nota. Este gráfico muestra todas las fases de la metodología de DSR. Adaptado de *Introduction to Design Science Research* (p. 6), por J. Brocke, A. Maedche y A. Hevner, 2020, Design Science Research Cases.

Identificación de la problemática. Esta actividad define el problema de investigación específico y justifica el valor de una solución (Brocke, Maedche, & Hevner, 2020). Para esta tarea se realizaron entrevistas y la revisión de literatura, esto con la finalidad de encontrar relación entre los problemas encontrados en el área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE.

Definición de los objetivos de la solución. Los objetivos de una solución se pueden inferir de la definición del problema y del conocimiento de lo que es posible y factible (Brocke, Maedche, & Hevner, 2020). En esta fase se realizó una matriz de congruencia metodología para determinar los objetivos específicos y sus respetivas preguntas de investigación que sirven para que los investigadores se guíen correctamente a la solución.

Diseño y desarrollo. Se crea un artefacto (Brocke, Maedche, & Hevner, 2020). En esta fase se diseñará la arquitectura de la solución con sus respectivas fases, además, se realizará el desarrollo del modelo machine learning.

Demostración. Esta actividad demuestra el uso del artefacto para resolver una o más instancias del problema (Brocke, Maedche, & Hevner, 2020). En esta fase se pondrá en funcionamiento el modelo machine learning y se probaran con distintos artículos científicos su comportamiento.

Evaluación. La evaluación mide qué tan bien el artefacto apoya una solución al problema (Brocke, Maedche, & Hevner, 2020). En este caso, para la evaluación del algoritmo se realizará una evaluación del rendimiento utilizando métricas como, por ejemplo, precisión, recall y puntuación F1.

Comunicación. Aquí todos los aspectos del problema y el artefacto diseñado se comunican a las partes interesadas relevantes (Brocke, Maedche, & Hevner, 2020). En este caso, junto con expertos del área de Biotecnología de la universidad se realizará una comparación entre los resultados obtenidos con el algoritmo y los datos extraídos manualmente.

Marco Teórico

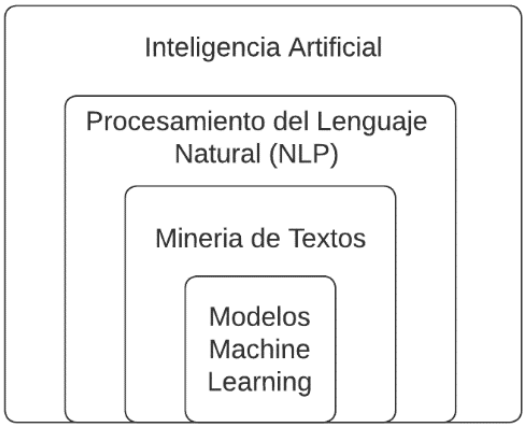
Red de categorías

Para encontrar una correcta consistencia en la parte teórica de este estudio se optó por la realización de una red de categorías para determinar las variables dependientes (VD) y las variables independientes (VI) (ver figura 10 y 11), cabe mencionar que esto se pudo determinar debido a la hipótesis que se planteó en el capítulo anterior donde se establecen las siguientes variables:

- **VI:** Modelo machine learning basado en técnicas de minería de textos y Procesamiento del Lenguaje Natural (NLP).
- **VD:** Automatizar tareas de recolección de datos en artículos científicos de Nanopartículas.

Figura 10

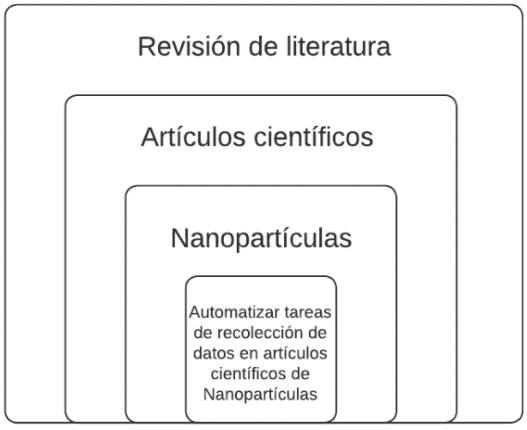
Red de categoría de la VI



Nota. Este gráfico muestra todos los temas a tratar dentro de la variable independiente.

Figura 11

Red de categoría de la VD



Nota. Este gráfico muestra todos los temas a tratar dentro de la variable dependiente.

Fundamentación Científica de la Variable Dependiente

Revisión de literatura. La revisión de literatura es una de las formas más sencillas de economizar esfuerzos en una investigación. Consiste en el repaso y reconstrucción de trabajos ya realizados por otros y tiene como fin el detectar, obtener y consultar la bibliografía y otros materiales que pueden ser útiles a los propósitos del estudio, así como extraer y recopilar la información relevante y necesaria que atañe a nuestro problema de investigación (Amador, 2009).

Las revisiones de literatura suele ser el primer paso antes de realizar una investigación ya que con ella nos aproximamos al conocimiento de un tema y ayuda a identificar lo que se sabe y lo que se desconoce del tema de interés.

Si se desea realizar una verdadera revisión integral de la literatura, el trabajo que se realice debe ofrecer al lector un resumen conciso, objetivo y lógico del conocimiento actual sobre un tema en particular (Guirao, 2015).

Artículos científicos. Un artículo científico presenta resultados de investigación escritos por investigadores y científicos. Por lo general, se consideran fuentes primarias y se escriben para otros investigadores. Los artículos más recientes contendrán el trabajo más reciente en el campo, con referencias a trabajos publicados anteriormente en el campo de estudio (Johnson, 2014).

Los artículos científicos generalmente contienen la siguiente estructura que incluye título, abstract, introducción, métodos, resultados, discusión, conclusiones y referencias.

El artículo científico tiene una serie de características, entre ellas, el hecho de que tiene que ser original, es decir, aportar algo nuevo al campo de estudio en el que se inserte la temática tratada. Los resultados que se presenten han de ser válidos y fidedignos, debe estar

escrito con un lenguaje claro y preciso y, sobre todo, utilizar una metodología con instrumentos y procedimientos que se haya demostrado que son científicamente válidos, independientemente de que en la investigación llevada a cabo se siga una metodología cuantitativa, cualitativa o mixta (Hernando, 2019).

Nanopartículas. Para esta investigación se utilizó el caso de estudio de las nanopartículas que sin duda es uno de los campos más estudiados dentro de la Biotecnología.

Las nanopartículas son en si partículas microscópicas con una dimensión menor a la de 100 nanómetros. Hoy en día el campo de las nanopartículas está experimentando una gran expansión en materia de investigación científica gracias a su potencial para ser usadas en sectores como la medicina, la electricidad, la cosmética o la óptica entre otras áreas (Solmeglas Lab, 2020).

Automatizar tareas de recolección de datos en artículos científicos de Nanopartículas. Cuando se realizan tareas de investigación, en este caso revisión de literatura, y se obtienen los artículos científicos esenciales sobre el estudio de nanopartículas, se extrae la información relevante dentro de estos como por ejemplo la familia, localidad, nombre, métodos, equipo, valores y demás datos importantes relacionados a dicha nanopartícula. Esto se hace con el propósito de experimentar, caracterizar, sintetizar, entre otras tareas sobre nanopartículas.

Además, la tarea de recolección de datos es muy importante ya que para producir una nanopartícula se necesitan condiciones definidas que en este caso se las encuentran en los artículos científicos y de esta manera se emplean procesos de síntesis específicos para producir diversos tipos de nanopartículas, recubrimientos, dispersiones o compuestos (Santos, 2022).

Fundamentación Científica de la Variable Independiente

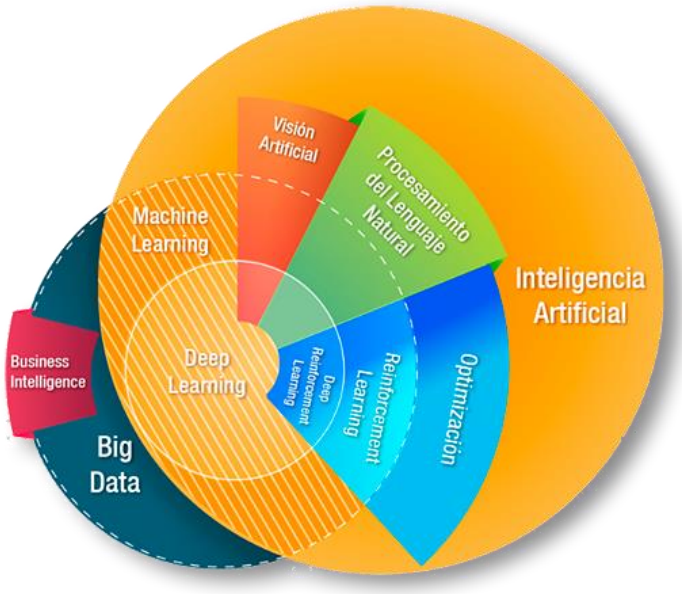
Inteligencia Artificial. La inteligencia artificial (IA) es una rama de gran alcance de las ciencias de la computación que se ocupa de construir máquinas inteligentes capaces de realizar tareas que normalmente requieren inteligencia humana (Builtin, 2022). Existen diferentes tipos de IA como se describe a continuación:

- **IA débil:** Es una IA entrenada y enfocada para realizar tareas específicas. La IA débil impulsa la mayor parte de la IA que nos rodea hoy. Permite algunas aplicaciones muy sólidas, como Siri de Apple, Alexa de Amazon, Watson de IBM y vehículos autónomos (IBM Cloud Education, 2020).
- **IA fuerte:** Esta compuesta por Inteligencia General Artificial (AGI) y Súper Inteligencia Artificial (ASI). La inteligencia artificial general (AGI), o IA general, es una forma teórica de IA en la que una máquina tendría una inteligencia igual a la de los humanos; tendría una conciencia autoconsciente que tiene la capacidad de resolver problemas, aprender y planificar para el futuro. La superinteligencia artificial (ASI), también conocida como superinteligencia, superaría la inteligencia y la capacidad del cerebro humano (IBM Cloud Education, 2020). Sin embargo, este tipo de IA se mantiene como teórica debido a que aún no existen ejemplos prácticos.

La IA es muy importante y por ello de esta parten diferentes campos de estudio como por ejemplo el Procesamiento del Lenguaje Natural (NLP), Machine Learning, Deep Learning, Visión artificial, entre otras como se puede ver en la figura 12.

Figura 12

Campos de la Inteligencia Artificial



Nota. Este gráfico muestra algunos campos que engloba la Inteligencia Artificial. Adaptado de *Big Data & Artificial Inteligence*, por ICC, 2022 (https://www.iic.uam.es/en/big-data-artificial-inteligence/https://www.iic.uam.es/en/big-data-artificial-inteligence/). Derechos de autor 2022 por ICC.

Existen numerosas aplicaciones en donde se puede aplicar la IA, entre las más comunes se tiene el reconocimiento de voz, los agentes de servicio al cliente, la visión por computadora, motores de recomendaciones, negociación de acciones automatizada, entre otros.

Sin duda, la IA es uno de los campos más estudiados en la actualidad ya que es una ciencia interdisciplinaria con múltiples enfoques, pero los avances en el machine learning y el deep learning están creando un cambio de paradigma en prácticamente todos los sectores de la industria tecnológica (Builtin, 2022).

Procesamiento del Lenguaje Natural (NLP). El Procesamiento del Lenguaje Natural (NLP) es el campo de conocimiento de la Inteligencia Artificial que se ocupa de investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino (Moreno, 2018).

El NLP toma elementos prestados de muchas disciplinas, incluyendo la ciencia de la computación y la lingüística computacional, en su afán por cerrar la brecha entre la comunicación humana y el entendimiento de las computadoras (SAS, 2022). Ya sea que el idioma sea hablado o escrito, el NLP utiliza inteligencia artificial para tomar información del mundo real, procesarla y darle sentido de una manera que una computadora pueda entender (Lutkevich, 2021).

NLP aplica varios algoritmos para comprender el significado y la estructura de las oraciones, por esta razón, usa técnicas basadas en la semántica, es decir, el contexto y significado de una palabra. A continuación, las técnicas semánticas:

- **Desambiguación del sentido de la palabra:** Se refiere al significado de una palabra en función de su contexto, por ejemplo, “the pi gis in the pen”, la palabra “pen” tiene varios significados, sin embargo, el algoritmo que use este método debe ser capaz de entender que en la oración no se hace referencia a un esfero, sino que hace referencia a un corral (Lutkevich, 2021).
- **Reconocimiento de entidades nombradas (NER):** Es el proceso de identificar grupos específicos de palabras que comparten características semánticas comunes. Esta es una subtarea de extracción de información que busca ubicar y clasificar entidades nombradas en el texto en categorías predefinidas, como los nombres de personas, organizaciones, ubicaciones, entre otras (Gupta, 2018).

- **Generación de lenguaje natural:** Esto utiliza una base de datos para determinar la semántica detrás de las palabras y generar texto nuevo (Lutkevich, 2021).

Con NLP se pueden realizar varias tareas como por ejemplo la categorización de contenido, descubrimiento y modelado de temas, extracción contextual de información, análisis de sentimientos, conversión de habla a texto y de texto a habla, resumen de documentos, traducción basada en máquina, entre otras.

Minería de textos. En la actualidad existe mucha información en formatos no estructurado como, por ejemplo, correos electrónicos, páginas web, archivos PDF, foros de noticias, etc. El recopilar esta información sería muy beneficioso para distintos análisis, debido a esta necesidad, la minería de textos aparece y ayuda a estos procesos de recopilación, exploración y aprovechamiento de toda esa información.

La minería de textos es el proceso de analizar colecciones de materiales de texto con el objeto de capturar los temas y conceptos clave y descubrir las relaciones ocultas y las tendencias existentes sin necesidad de conocer las palabras o los términos exactos que los autores han utilizado para expresar dichos conceptos (IBM, 2021).

Modelos Machine Learning. Antes de conocer los modelos machine learning, es importante comprender su definición. Como se vio anteriormente, el machine learning es una rama de la IA que se encarga de crear sistemas que aprenden automáticamente. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros (González, 2022).

A medida que los algoritmos ingieren los datos de entrenamiento, se pueden generar modelos más precisos basados en esos datos. Un modelo de machine learning es la salida que se genera cuando entrena su algoritmo de machine learning con datos. Después del

entrenamiento, cuando proporcione una entrada a un modelo, se le dará una salida (IBM, 2022).

Dentro del machine learning existen varios enfoques para los entrenamientos de los modelos, a continuación, se los menciona:

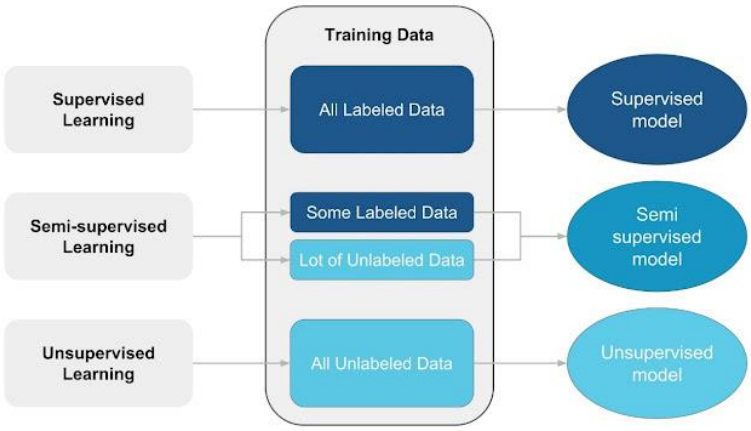
Aprendizaje supervisado: El aprendizaje supervisado normalmente comienza con un conjunto establecido de datos y un determinado nivel de comprensión sobre cómo se clasifican los datos (IBM, 2022). Este tipo de aprendizaje trata de encontrar patrones en los datos y estos datos tienen que estar etiquetados para de esta manear definir su significado, es por esta razón que para este proyecto se optó por el uso de este enfoque.

Aprendizaje no supervisado: El aprendizaje no supervisado se utiliza cuando el problema requiere una gran cantidad de datos sin etiquetar (IBM, 2022). Este tipo de aprendizaje realiza un proceso iterativo para analizar los datos sin que intervengan humanos.

Aprendizaje semi supervisado: El aprendizaje semi supervisado a diferencia de los dos enfoques vistos anteriormente, utiliza pocos datos etiquetados y muchos datos no etiquetados como parte de su conjunto de entrenamiento, este enfoque trata de explorar la información estructural que contienen los datos no etiquetados con el objetivo de generar modelos predictivos que funcionen mejor que los que sólo utilizan datos etiquetados (Ibáñez, 2019). En la figura 13 se pueden visualizar los enfoques del machine learninig.

Figura 13

Enfoques en el Machine Learning



Nota. Este gráfico muestra los enfoques que componen al machine learning. Adaptado de *Semi-Supervised Learning…el gran desconocido*, por A. Ibáñez, 2019 (https://empresas.blogthinkbig.com/semi-supervised-learningel-gran-desconocido/). Derechos de autor 2019 por Telefónica Tech.

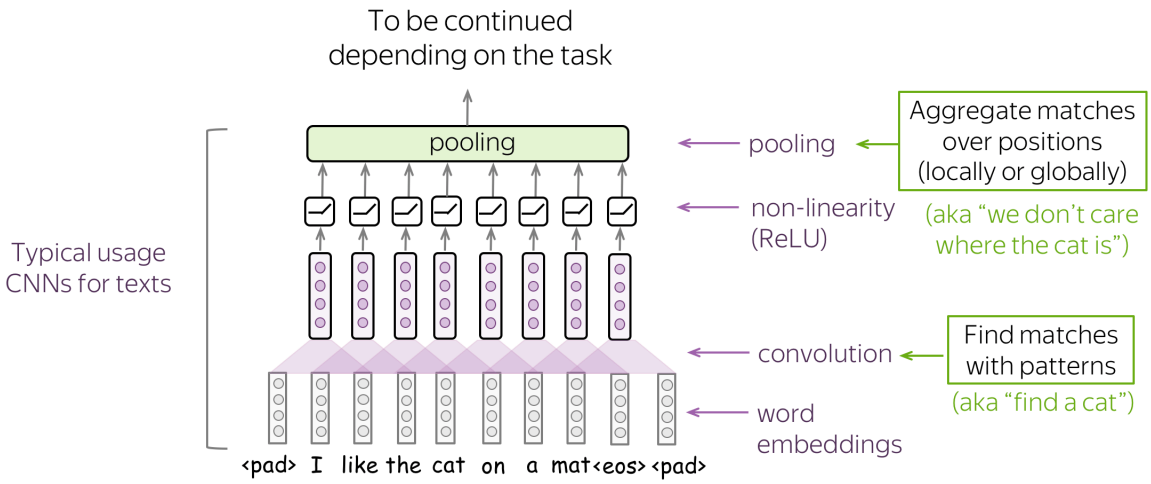
Una vez comprendido la definición y todo lo que conlleva al machine learning se procederá a la explicación de los modelos.

Redes Neuronales Convolucionales para texto: Para ello se parte de la definición de una red neuronal convolucional normal que en este caso son un tipo de redes neuronales artificiales donde las neuronas corresponden a campos receptivos de una manera muy similar a las neuronas en la corteza visual primaria (V1) de un cerebro biológico. Este tipo de red es una variación de un perceptrón multicapa, sin embargo, debido a que su aplicación es realizada en matrices bidimensionales, son muy efectivas para tareas de visión artificial, como en la clasificación y segmentación de imágenes, entre otras aplicaciones (Barrios, 2022).

Este tipo de redes suele ser muy utilizada para imágenes, sin embargo, también es posible aplicarlas dentro de textos ya estos tendrían una sola dimensión por lo tanto las convoluciones serian unidimensionales.

Figura 14

Red Neuronal Convolucional para textos



Nota. Este gráfico muestra los enfoques que componen al machine learning. Adaptado de *Convolutional Neural Networks for Text*, por L. Voita, 2022 (https://lena-voita.github.io/nlp_course/models/convolutional.html). Derechos de autor 2022 por Lena Voita.

En la figura 14 se muestra un modelo convolucional típico para textos. Por lo general, se aplica una capa convolucional a la incrustación de palabras, seguida de una no linealidad (generalmente ReLU) y una operación de agrupación. Estos son los principales componentes básicos de los modelos convolucionales: para tareas específicas, las configuraciones pueden ser diferentes, pero estos bloques son estándar (Voita, 2022).

BERT: BERT es un acrónimo de Representaciones de codificador bidireccional de Transformer. Eso significa que, a diferencia de la mayoría de las técnicas que analizan oraciones de izquierda a derecha o de derecha a izquierda, BERT va en ambas direcciones usando el codificador Transformer (Cardellino, 2021). BERT es una técnica de NLP que fue creado por Google en 2018 y es una biblioteca de código abierto.

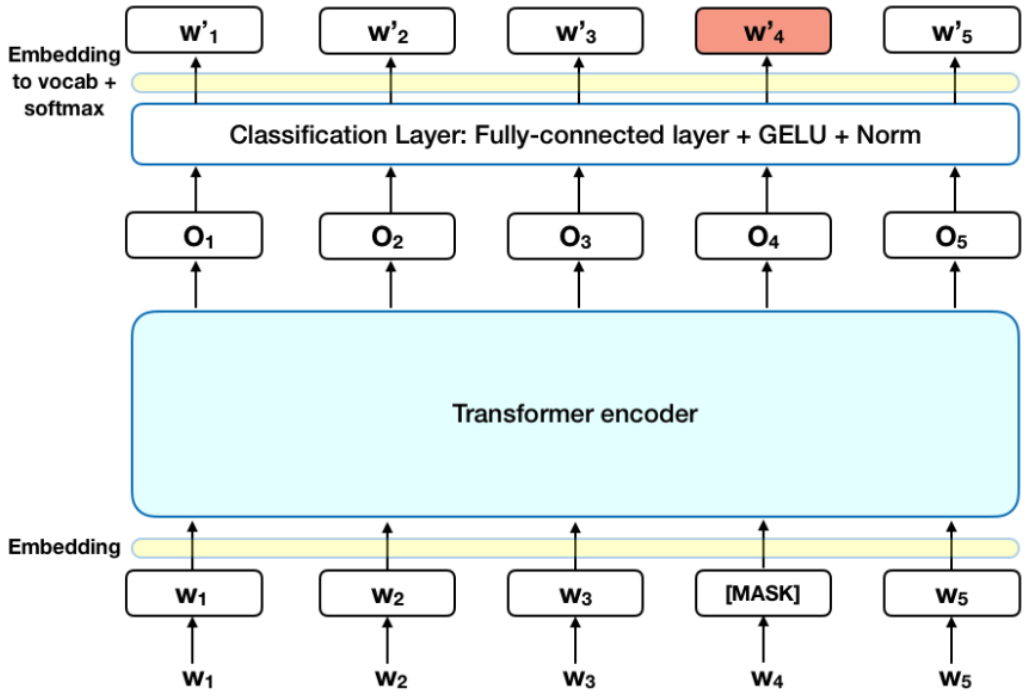
Este modelo brinda resultados muy precisos y con rendimientos bantante buenos, sin embargo, para llegar a este nivel se necesita de una gran cantidad de datos etiquetados.

El enfoque bidireccional que utiliza significa que obtiene más contexto para una palabra que si solo estuviera entrenando en una dirección. Con este contexto adicional, puede aprovechar otra técnica llamada LM enmascarada (Cardellino, 2021).

De manera resumida, para que funciones BERT se necesita agregar una capa de clasificación encima de la salida del codificador, después se multiplican los vectores de salida por la matriz de incrustación, transformándolos en la dimensión del vocabulario, finalmente se calcula la probabilidad de cada palabra en el vocabulario con softmax. En la figura 15 se puede visualizar la arquitectura de BERT.

Figura 15

Arquitectura de BERT



Nota. Este gráfico muestra las capas que componen la arquitectura de BERT. Adaptado de *BERT Explained: State of the art language model for NLP*, por R. Horev, 2018 (https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270)). Derechos de autor 2018 por Towards Data Science.

RoBERTa: Se podría definir a RoBERTa (Enfoque de preentrenamiento BERT robustamente optimizado) como una optimización mejorada de las Representaciones de codificador bidireccional de Transformers o mejor conocido como BERT, desarrollado por Facebook en 2019.

RoBERTa se basa en la estrategia de enmascaramiento de lenguaje de BERT, en la que el sistema aprende a predecir secciones de texto intencionalmente ocultas dentro de ejemplos de lenguaje sin anotaciones. RoBERTa, que se implementó en PyTorch, modifica los hiperparámetros clave en BERT, incluida la eliminación del objetivo de entrenamiento previo de la siguiente oración de BERT y el entrenamiento con mini lotes y tasas de aprendizaje mucho más grandes (Meta AI, 2019).

Capítulo III

En esta sección se presenta el proceso de desarrollo del modelo machine learning basado en el caso de estudio de nanopartículas para la extracción automática de información en artículos científicos que responda la necesidad actual del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE.

Arquitectura de la solución

La arquitectura utilizada para la creación de la solución de este proyecto comprende las siguientes cuatro fases (ver figura 16).

Primero, la fase de **Origen de los datos**, en esta fase se identificaron las variables que son necesarias dentro de los artículos científicos, después se realizó la recolección de datos que consiste en la obtención de varios artículos científicos sobre nanopartículas y, además, la elaboración de una pequeña base de datos en Excel que contiene la extracción manual de la información de dichos artículos científicos.

Después, se realizó un análisis de todos los datos para descartar variables innecesarias dentro del estudio. Esta fase es una de las más importantes puesto que, al ser un caso de estudio sobre nanopartículas se necesita personal experto en este campo para recolectar datos válidos, debido a lo antes mencionado, esta fase fue desarrollada por personal profesional del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE.

Segundo, la fase de **Proceso ETL** (cuyas siglas significan Extraer-Transformar-Cargar) este proceso hace referencia a mover datos de una o más fuentes, realizar algunos cambios y luego cargarlos en un nuevo destino único (Mahmud, 2021).

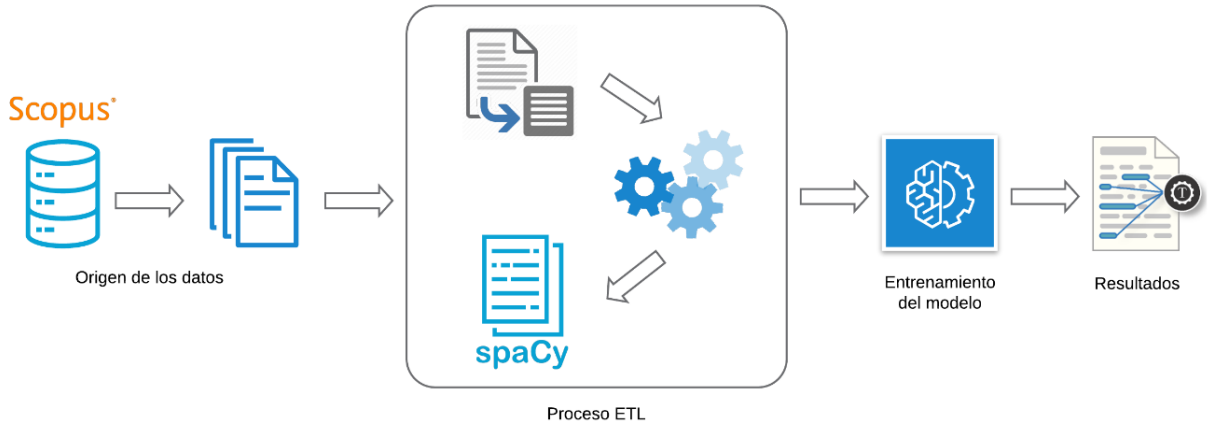
En la etapa de “Extracción” se limpió y extrajo los textos de los artículos científicos que eran archivos no estructurados, es decir, archivos PDF; en la etapa de “Transformación” se

realizó el proceso de tokenizado[3], etiquetamiento de los datos y limpieza; finalmente en la etapa de “Carga” se adecuo los datos para que cumplan con el formato requerido en la fase de entrenamiento del modelo.

Tercero, la fase de **Entrenamiento del modelo** se utilizó la librería spaCy de Python ya que cuenta con algoritmos de Procesamiento del Lenguaje Natural (NLP) y, además, permite construir modelos NER.

Figura 16

Arquitectura de la solución



Nota. Este gráfico detalla la arquitectura de la solución y sus respectivas fases.

Finalmente, en la fase de **Resultados** de igual manera se utilizó la librería spaCy de Python ya que cuenta con comandos para evaluar el rendimiento del modelo. Esta fase se presenta con más detalle en el siguiente capítulo.

[3] Tokenizado: Separar palabras del texto en entidades llamadas tokens.

Fases de la arquitectura

Origen de los datos

Como se mencionó anteriormente en esta fase se realizó un estudio de las variables necesarias para que a partir de estas realizar la búsqueda de artículos científicos sobre nanopartículas en la base de datos digital Scopus, cabe mencionar que solo se utilizaron artículos en el idioma inglés ya que, si se mezclan idiomas, como el español, haría que el modelo machine learning sea impreciso.

Para este proyecto se logró recaudar un total de 100 artículos científicos que fueron sometidos a un estudio individual. Además, esta fue una de las fases más laboriosas tomando en cuenta que se obtuvo ayuda de dos personas del área de Biotecnología.

Comprensión de los datos. Según las necesidades del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE, identificaron un total de 141 entidades las cuales se utilizaron para realizar posteriormente una base de datos en el programa de Excel.

Recolección de datos. Una identificadas las entidades, se utilizó la base de datos Scopus donde se recolectó 100 artículos científicos que cumplían con las características necesarias para posteriormente llenar la base de datos en Excel manualmente. En la tabla 6 se detalla el material entregado por parte del personal de biotecnología.

Tabla 6

Material entregado

Material	Cantidad	Detalle
Artículos sin anotaciones	100	Artículos descargados directamente de la base de datos Scopus sin ninguna modificación.
Artículos con anotaciones	100	Artículos donde se encuentra subrayado la información necesaria a extraer.
Base de datos	1	Base de datos en Excel con la información extraída de los artículos, esta información se puede comparar con los artículos subrayados.

Nota. Esta tabla muestra el material entregado por parte del personal de Biotecnología.

Análisis de los datos. Una vez con el material, se procedió a realizar un análisis minucioso de toda la base de datos y de igual manera de los artículos científicos entregados. Una vez realizado el análisis se concluyó que, de las 141 entidades identificadas, para el estudio eran viables 45 entidades, esto debido a que las entidades eliminadas no contenían suficiente información. Además, de los 100 artículos entregados, 88 fueron utilizados para el estudio, ya que los 12 artículos restantes contenían marcas de agua, números en los costados, entre otros. A continuación, se muestran los artículos seleccionados (ver tabla 7).

Tabla 7

Artículos seleccionados

N	R	Titulo
1	X9	Characterization of silver nanoparticles by green synthesis method using Pedalium murex leaf extract and their antibacterial activity
2	X11	Green synthesis of silver nanoparticles from leaf extract of Mimusops elengi, Linn. for enhanced antibacterial activity against multi drug resistant clinical isolates
3	X13	Leaf extract mediated green synthesis of silver nanoparticles from widely available Indian plants: Synthesis, characterization, antimicrobial property and toxicity analysis
4	X18	Green biosynthesis and characterization of magnetic iron oxide (Fe 3O4) nanoparticles using seaweed (Sargassum muticum) aqueous extract
5	X19	Biosynthesis of zinc oxide nanoparticles from Azadirachta indica for antibacterial and photocatalytic applications
6	X26	Green synthesis of Fe nanoparticles using eucalyptus leaf extracts for treatment of eutrophic wastewater
7	X28	Green synthesis of ZnO nanoparticles using Solanum nigrum leaf extract and their antibacterial activity

N	R	Título
8	X33	Biosynthesis, characterization and antimicrobial activity of copper oxide nanoparticles (CONPs) produced using brown alga extract (Bifurcaria bifurcata)
9	X35	Green biosynthesis and characterization of zinc oxide nanoparticles using brown marine macroalga Sargassum muticum aqueous extract
10	X36	Origanum vulgare mediated biosynthesis of silver nanoparticles for its antibacterial and anticancer activity
11	X41	Green synthesis of colloidal copper oxide nanoparticles using Carica papaya and its application in photocatalytic dye degradation
12	X43	Facile green synthesis of silver nanoparticles using Berberis vulgaris leaf and root aqueous extract and its antibacterial activity
13	X47	Biosynthesis, characterisation and anti-bacterial effect of plant-mediated silver nanoparticles using Artemisia nilagirica
14	X53	Green synthesis and characterization of silver nanoparticles using Boerhaavia diffusa plant extract and their anti bacterial activity
15	X54	Coleus aromaticus leaf extract mediated synthesis of silver nanoparticles and its bactericidal activity
16	X56	Sesbania grandiflora leaf extract mediated green synthesis of antibacterial silver nanoparticles against selected human pathogens

N	R	Título
17	X58	ZnO nanoparticles via Moringa oleifera green synthesis: Physical properties & mechanism of formation
18	X59	Green synthesis of gold nanoparticles using Citrus fruits (Citrus limon, Citrus reticulata and Citrus sinensis) aqueous extract and its characterization
19	X65	A study on the stability and green synthesis of silver nanoparticles using Ziziphora tenuior (Zt) extract at room temperature
20	X67	Evaluation of antioxidant, antibacterial and cytotoxic effects of green synthesized silver nanoparticles by Piper longum fruit
21	X69	Biobased green method to synthesise palladium and iron nanoparticles using Terminalia chebula aqueous extract
22	X73	Green synthesis of titanium dioxide nanoparticles using Psidium guajava extract and its antibacterial and antioxidant properties
23	X82	Anticancer activity of Ficus religiosa engineered copper oxide nanoparticles
24	X83	Synthesis of monodispersed silver nanoparticles using Hibiscus cannabinus leaf extract and its antimicrobial activity
25	X86	Biosynthesis of silver nanoparticles using Alternanthera sessilis (Linn.) extract and their antimicrobial, antioxidant activities

N	R	Título
26	X94	Synthesis of cerium oxide nanoparticles using Gloriosa superba L. leaf extract and their structural, optical and antibacterial properties
27	X97	Antioxidant and anti-inflammatory activities of zinc oxide nanoparticles synthesized using Polygala tenuifolia root extract
28	X102	Green synthesis of silver nanoparticles using Alternanthera dentata leaf extract at room temperature and their antimicrobial activity
29	X105	Green synthesis of palladium nanoparticles using Hippophae rhamnoides Linn leaf extract and their catalytic activity for the Suzuki-Miyaura coupling in water
30	X107	Aloe vera extract functionalized zinc oxide nanoparticles as nanoantibiotics against multi-drug resistant clinical bacterial isolates
31	X108	Bark extract mediated green synthesis of silver nanoparticles: Evaluation of antimicrobial activity and antiproliferative response against osteosarcoma
32	X113	Green synthesis of mesoporous hematite (α-Fe2O 3) nanoparticles and their photocatalytic activity
33	X115	Green Synthesis of Magnetite (Fe3O4) Nanoparticles Using Seaweed (Kappaphycus alvarezii) Extract

N	R	Título
34	X116	Green synthesis of silver nanoparticles from marigold flower and its synergistic antimicrobial potential
35	X124	Synthesis of silver nanoparticles using reducing agents obtained from natural sources (Rumex hymenosepalus extracts)
36	X127	One-step green synthesis and characterization of leaf extract-mediated biocompatible silver and gold nanoparticles from Memecylon umbellatum
37	X128	Ecofriendly synthesis of silver nanoparticles from commercially available plant powders and their antibacterial properties
38	X129	Green synthesis, antimicrobial and cytotoxic effects of silver nanoparticles using Eucalyptus chapmaniana leaves extract
39	X131	Green synthesis and applications of Au-Ag bimetallic nanoparticles
40	X133	Green biosynthesis of silver nanoparticles using Calliandra haematocephala leaf extract, their antibacterial activity and hydrogen peroxide sensing capability
41	X135	Euphorbia heterophylla leaf extract mediated green synthesis of Ag/TiO2 nanocomposite and investigation of its excellent catalytic activity for reduction of variety of dyes in water

N	R	Título
42	X136	Mosquitocidal and antibacterial activity of green-synthesized silver nanoparticles from Aloe vera extracts: towards an effective tool against the malaria vector Anopheles stephensi?
43	X137	Green synthesis of silver nanoparticles using polysaccharides extracted from marine macro algae
44	X139	Biosynthesis and characterization of Acalypha indica mediated copper oxide nanoparticles and evaluation of its antimicrobial and anticancer activity
45	X140	Synthesis of iron-based nanoparticles using oolong tea extract for the degradation of malachite green
46	X151	Green Synthesis of Copper Oxide Nanoparticles Using Aloe vera Leaf Extract and Its Antibacterial Activity Against Fish Bacterial Pathogens
47	X155	Green synthesis of Co3O4 nanoparticles via Aspalathus linearis: Physical properties
48	X156	Green biosynthesis of gold nanoparticles using Galaxaura elongata and characterization of their antibacterial activity
49	X158	Green synthesis of silver nanoparticles from the extract of the inflorescence of Cocos nucifera (Family: Arecaceae) for enhanced antibacterial activity

N	R	Título
50	X167	Green synthesis of silver nanoparticles using marine algae Caulerpa racemosa and their antibacterial activity against some human pathogens
51	X173	Microwave-assisted green synthesis of silver nanoparticles using orange peel extract
52	X176	Green synthesis of silver nanoparticles mediated by Pulicaria glutinosa extract
53	X180	Green synthesis and characterization of silver nanoparticles using Lantana camara leaf extract
54	X183	The Green synthesis of gold nanoparticles using an aqueous root extract of Morinda citrifolia L
55	X188	Biofabrication of zinc oxide nanoparticles using fruit extract of Rosa canina and their toxic potential against bacteria: A mechanistic approach
56	X198	Characterization of iron-polyphenol nanoparticles synthesized by three plant extracts and their fenton oxidation of azo dye
57	X201	Green synthesis of copper nanoparticles by Citrus medica Linn. (Idilimbu) juice and its antimicrobial activity
58	X207	Green synthesis of copper oxide nanoparticles using abutilon indicum leaf extract: Antimicrobial, antioxidant and photocatalytic dye degradation activities

N	R	Título
59	X217	Antibacterial and cytotoxic effect of biologically synthesized silver nanoparticles using aqueous root extract of Erythrina indica lam
60	X228	Green synthesis of silver nanoparticles, their characterization, application and antibacterial activity
61	X229	Greener approach for synthesis of antibacterial silver nanoparticles using aqueous solution of neem gum (Azadirachta indica L.)
62	X241	Green synthesis of silver and gold nanoparticles using Zingiber officinale root extract and antibacterial activity of silver nanoparticles against food pathogens
63	X252	Biomolecule-mediated synthesis of selenium nanoparticles using dried vitis vinifera (raisin) extract
64	X255	Green synthesis of silk fibroin-silver nanoparticle composites with effective antibacterial and biofilm-disrupting properties
65	X262	Biogenic nano-scale silver particles by Tephrosia purpurea leaf extract and their inborn antimicrobial activity
66	X263	Green synthesis of silver nanoparticles using Pinus eldarica bark extract
67	X276	A549 lung cell line activity of biosynthesized silver nanoparticles using Albizia adianthifolia leaf

N	R	Título
68	X279	Green synthesis, characterization and antimicrobial activity of Au NPs using Euphorbia hirta L. leaf extract
69	X293	Biogenic copper oxide nanoparticles synthesis using Tabernaemontana divaricate leaf extract and its antibacterial activity against urinary tract pathogen
70	X296	Green synthesis of ZnO nanoparticles by Aspalathus linearis: Structural & optical properties
71	X308	Synthesis of ecofriendly copper oxide nanoparticles for fabrication over textile fabrics: Characterization of antibacterial activity and dye degradation potential
72	X309	Biogenesis of copper oxide nanoparticles (CuONPs) using Sida acuta and their incorporation over cotton fabrics to prevent the pathogenicity of Gram negative and Gram positive bacteria
73	X317	Green synthesis of titanium dioxide (TiO2) nanoparticles by Trigonella foenum-graecum extract and its antimicrobial properties
74	X319	Biosynthesis of silver nanoparticles from Aloe vera leaf extract and antifungal activity against Rhizopus sp. and Aspergillus sp.
75	X331	Green synthesis of SnO2 nanoparticles and its photocatalytic activity of phenolsulfonphthalein dye

N	R	Título
76	X342	Green synthesis of gold nanoparticles and their anticancer activity
77	X354	Characterization and biotoxicity of Hypnea musciformis-synthesized silver nanoparticles as potential eco-friendly control tool against Aedes aegypti and Plutella xylostella
78	X386	Antimicrobial activities of silver nanoparticles synthesized from Lycopersicon esculentum extract
79	X388	Facile green synthesis of variable metallic gold nanoparticle using Padina gymnospora, a brown marine macroalga
80	X395	Green synthesis of gold nanoparticles using brown algae Cystoseira baccata: Its activity in colon cancer cells
81	X398	Green synthesis and spectral characterization of silver nanoparticles from Lakshmi tulasi (Ocimum sanctum) leaf extract
82	X407	Green synthesis of ZnO nanoparticles and its application in the degradation of some dyes
83	X421	Green synthesis of silver nanoparticles using marine macroalga Chaetomorpha linum
84	X423	Green synthesis of Monteponite CdO nanoparticles by Agathosma betulina natural extract

N	R	Título
85	X426	Green synthesis of silver nanoparticles by Trichoderma harzianum and their bio-efficacy evaluation against Staphylococcus aureus and Klebsiella pneumonia
86	X430	Green synthesis of iron oxide nanoparticles by aqueous leaf extract of Daphne mezereum as a novel dye removing material
87	X436	Green synthesis of silver nanoparticles using Rheum palmatum root extract and their antibacterial activity against Staphylococcus aureus and Pseudomonas aeruginosa
88	X440	Rosmarinus officinalis leaf extract mediated green synthesis of silver nanoparticles and investigation of its antimicrobial properties

Nota. Esta tabla muestra los artículos científicos que fueron seleccionados para el proyecto.

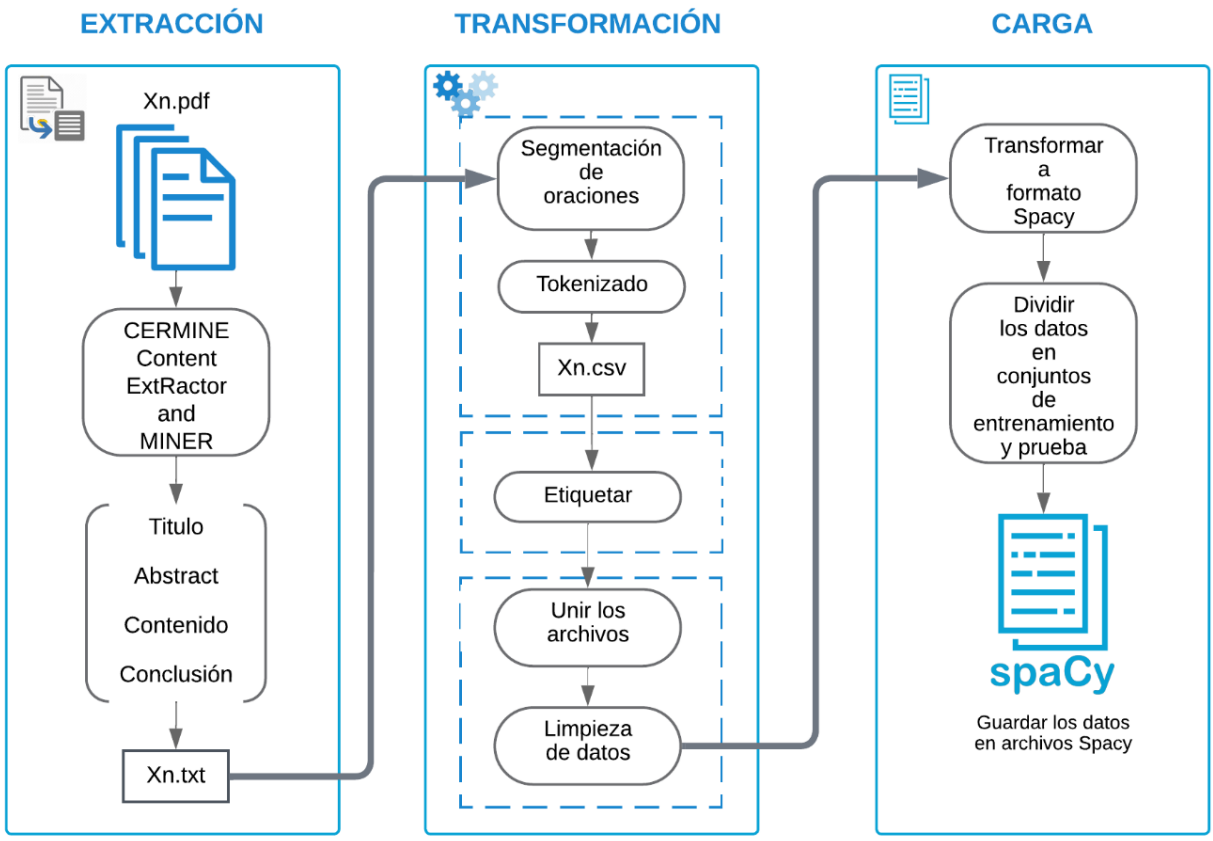
Proceso ETL

Como se mencionó anteriormente en esta fase se realizó un proceso sistemático denominado ETL para poder transformar los artículos científicos en datos capaces de ser entendidos por el modelo.

Además, esta es la fase más importante ya que si no se la realiza correctamente no será posible entrenar y posteriormente utilizar el modelo machine learning. Por consiguiente, esta etapa fue la más laboriosa. A continuación, se mostrará la arquitectura de este proceso.

Figura 17

Arquitectura del proceso ETL



Nota. Este gráfico detalla la arquitectura detallada del proceso ETL.

Extracción. En esta etapa, al ser los artículos científicos en formato PDF, la extracción de texto se convierte en una tarea compleja debido a que los artículos científicos están compuestos por muchas secciones y que, además, en su mayoría están en un formato de dos columnas. Por esta razón se optó por utilizar la herramienta CERMINE (Content ExtRactor and MINEr) de código abierto. A continuación, se muestra el uso de dicha herramienta:

- Subir el archivo PDF en su página web (http://cermine.ceon.pl/cermine/index.html)

- Después de esperar unos segundos la página desplegará los metadatos y el texto completo de dicho documento.

Figura 18

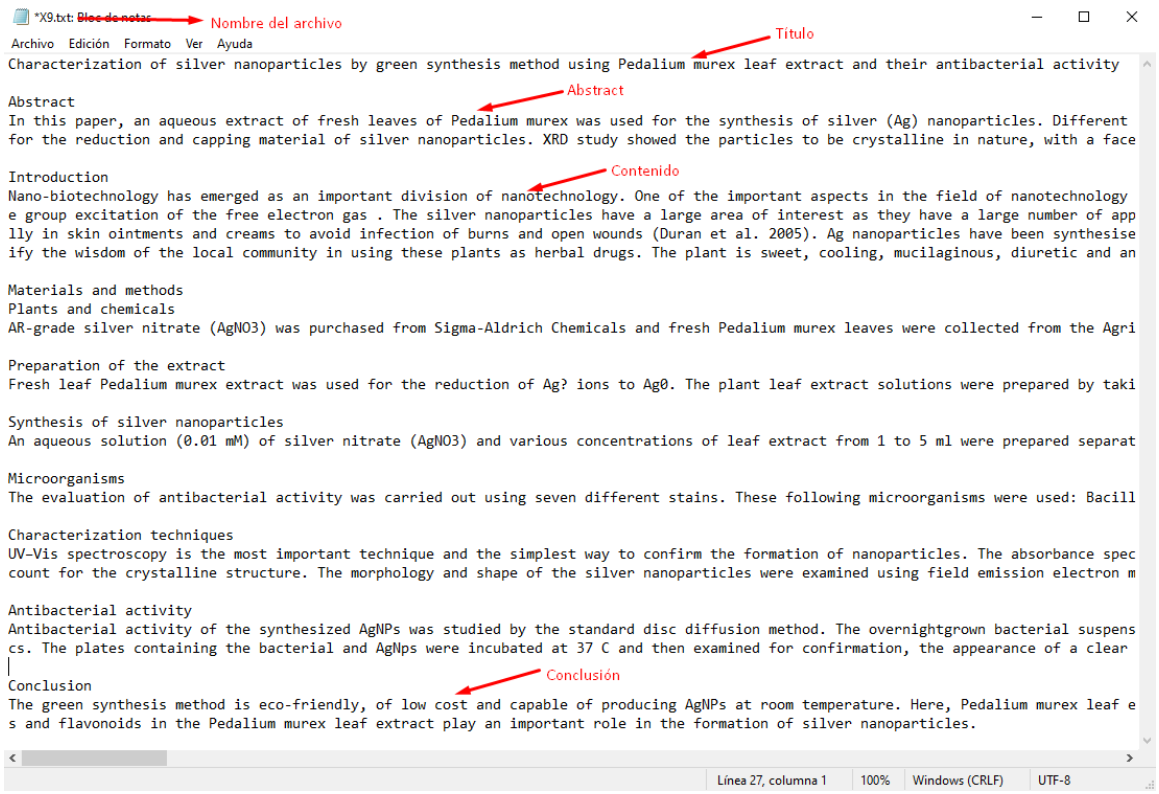
Pasos para utilizar CERMINE

Nota. Este gráfico muestra los pasos necesarios para utilizar la herramienta CERMINE.

Una vez se obtiene la información por parte de la herramienta se debe crear un archivo de texto (TXT) donde se pegará el Título, Abstract, Contenido y Conclusiones. Después de esto se tiene que hacer una limpieza manual ya que si bien, la herramienta ayuda a extraer la mayor parte del texto de manera exitosa siempre queda información basura como las descripciones de tablas e imágenes, por lo tanto, para tener una mejor calidad de datos es necesario realizar este paso. El proceso se realizó con todos los artículos científicos seleccionados y fueron guardados con el mismo nombre al que hacen referencia (Xn.txt). A continuación, se muestra el formato que tienen los archivos de texto:

Figura 19

Formato de los archivos de texto



Nota. Este gráfico muestra como tienen que guardarse los archivos de texto de cada artículo científico.

Transformación. Una vez extraído los textos de los artículos científicos a archivos de texto plano (TXT) se procede a realizar una serie de tareas que son necesarias para poder obtener el formato requerido para el entrenamiento del modelo. A continuación, se detallan las tareas que se realizaron.

Para los procesos detallados a continuación se hizo uso de las librerías de Python “Numpy”, “Pandas”, “NLTK”, “Cleantext”.

Segmentación de oraciones. Proceso que divide el texto en oraciones. Para este proceso se utilizó la librería de “NLTK” específicamente el método “sent_tokenize”.

Tokenizado. Proceso que divide las oraciones en palabras que son denominadas tokens. Para este proceso se utilizó la librería de “NLTK” específicamente el método “word_tokenize”.

- **Entrada:** “Characterization of silver nanoparticles”
- **Salida:** “[‘Characterization’, ‘of’, ‘silver’, ‘nanoparticles’]”

Después, se procedió a guardar cada token en archivos de extensión CSV (valores separados por coma) donde cada archivo contenía los siguiente (ver figura 20):

- **id_paper**: El nombre del archivo de texto.
- **id_sentence**: El identificados de cada oración que va desde 0 a n.
- **words**: Los tokens extraídos de los archivos de texto.

Figura 20

Contenido de los archivos CSV

id_paper	id_sentence	words
X9.txt	0	characterization
X9.txt	0	of
X9.txt	0	silver
X9.txt	0	nanoparticles
X9.txt	0	by
X9.txt	0	green
X9.txt	0	synthesis
X9.txt	0	method
X9.txt	0	using
X9.txt	0	pedalium
X9.txt	0	murex
X9.txt	0	leaf
X9.txt	0	extract
X9.txt	0	and
X9.txt	0	their
X9.txt	0	antibacterial
X9.txt	0	activity
X9.txt	1	abstract

Nota. Este gráfico muestra como tienen que guardarse los archivos CSV de cada artículo científico.

Etiquetar. Una vez obtenidos los tokens de cada artículo científico se procedió con el etiquetamiento de los datos, para esta tarea se utilizó el método mencionado en el Estado del Arte que es el uso de “BIO encoding” (Beginning-Inside-Outside), este es un formato común para etiquetar tokens. Antes de mostrar el proceso de etiquetado, a continuación, se muestran las 45 entidades que se utilizaron en este proyecto.

Tabla 8

Entidades

N	Descripción	Entidad	Etiqueta
1	Agente Reductor	Tipo	**AIA_TIPO**
2	Agente Reductor	Nombre Común	**AIA_NOMBRE**
3	Agente Reductor	Especie	**AIA_ESPECIE**
4	Agente Reductor	Familia	**AIA_FAMILIA**
5	Agente Reductor	Parte	**AIA_PARTE**
6	Agente Reductor	Localidad	**AIA_LOCALIDAD**
7	Agente Reductor	Propiedades	**AIA_PROPIEDAD**
8	Agente Reductor	Recolección	**AMP_RECOLECCION**
9	Agente Reductor	Agente de Lavado	**AMP_AGELAV**
10	Agente Reductor	Temperatura Secado	**AMP_TEMPSEC**
11	Agente Reductor	Técnica de reducción de tamaño	**AMP_TECREDTAM**

N	Descripción	Entidad	Etiqueta
12	Agente Reductor	Reducción de tamaño	**AMP_REDTAM**
13	Agente Reductor	Cantidad	**AMP_CANTIDAD**
14	Agente Reductor	Solvente	**AMP_SOLVENTE**
15	Agente Reductor	Volumen Solvente	**AMP_VOLSOL**
16	Agente Reductor	Técnica	**ATP_TECNICA**
17	Agente Reductor	Tiempo	**ATP_TIEMPO**
18	Agente Reductor	Temperatura	**ATP_TEMPERATURA**
19	Agente Reductor	Equipo	**ATP_EQUIPO**
20	Agente Reductor	Técnica	**ACA_TECNICA**
21	Agente Reductor	Variable	**ACA_VARIABLE**
22	Agente Reductor	Valor	**ACA_VALOR**
23	Agente Reductor	Interpretación	**ACA_INTERPRE**
24	Proceso De Síntesis De Nanopartículas	Agente Reductor (AR)	**PSN_AR**
25	Proceso De Síntesis De Nanopartículas	Volumen AR	**PSN_VOLAR**

N	Descripción	Entidad	Etiqueta
26	Proceso De Síntesis De Nanopartículas	Sal Precursora (SP)	**PSN_SALPRE**
27	Proceso De Síntesis De Nanopartículas	Volumen SP	**PSN_VOLSP**
28	Proceso De Síntesis De Nanopartículas	Concentración SP	**PSN_CONSP**
29	Proceso De Síntesis De Nanopartículas	Técnica	**PCM_TECNICA**
30	Proceso De Síntesis De Nanopartículas	Equipo	**PCM_EQUIPO**
31	Proceso De Síntesis De Nanopartículas	Temperatura	**PCM_TEMPERATURA**
32	Proceso De Síntesis De Nanopartículas	Tiempo	**PCM_TIEMPO**
33	Proceso De Síntesis De Nanopartículas	Tipo De NPs Obtenidas	**PSN_TIPNPSOBT**
34	Caracterización De Nanopartículas	Técnica	**CN_TECNICA**

N	Descripción	Entidad	Etiqueta
35	Caracterización De Nanopartículas	Variable	**CN_VARIABLE**
36	Caracterización De Nanopartículas	Valor	**CN_VALOR**
37	Caracterización De Nanopartículas	Interpretación	**CN_INTERPRETACION**
38	Caracterización De Nanopartículas	Equipo	**CN_EQUIPO**
39	Aplicación	"Actividad / Capacidad	**AP_ACTCAP**
40	Aplicación	Método	**AMC_METODO**
41	Aplicación	Microorganismo	**AMC_MICROO**
42	Aplicación	Medio	**AMC_MEDIO**
43	Aplicación	Método	**ACM_METODO**
44	Aplicación	Temperatura	**ACM_TEMPERATURA**
45	Aplicación	Tiempo	**ACM_TIEMPO**

Nota. Esta tabla muestra las entidades que se utilizaron para etiquetar a los tokens.

Una vez identificado las etiquetas que tendrá cada entidad identificada se procede a etiquetar los 88 archivos CSV manualmente, además se agregó una columna con el nombre de “tag”, en esta columna se etiqueto cada token con su respectiva etiqueta (ver figura 21).

Cabe mencionar que cada etiqueta debe estar acompañada con su respectivo prefijo, es decir, el prefijo B- significa el comienzo de un fragmento, el prefijo I- significa que la etiqueta está dentro de un fragmento y la etiqueta O significa que el token no pertenece a ninguna entidad (Kuriakose, 2019).

Figura 21

Contenido de los archivos CSV etiquetados

id_paper	id_sentence	words	tag
X9.txt	0	characterizat	O
X9.txt	0	of	O
X9.txt	0	silver	B-PSN_TIPNPSOBT
X9.txt	0	nanoparticle	O
X9.txt	0	by	O
X9.txt	0	green	O
X9.txt	0	synthesis	O
X9.txt	0	method	O
X9.txt	0	using	O
X9.txt	0	pedalium	B-AIA_ESPECIE
X9.txt	0	murex	I-AIA_ESPECIE
X9.txt	0	leaf	B-AIA_PARTE
X9.txt	0	extract	B-PSN_AR
X9.txt	0	and	O
X9.txt	0	their	O
X9.txt	0	antibacterial	B-AP_ACTCAP
X9.txt	0	activity	I-AP_ACTCAP
X9.txt	1	abstract	O
X9.txt	2	in	O

Nota. Este gráfico muestra un ejemplo de cómo están etiquetados los archivos.

En cuanto a la calidad de etiquetado de los datos, este proceso se realizó de acuerdo con el material entregado por el área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE, es decir, se siguió los artículos científicos subrayados y la base de datos en Excel con los datos extraídos manualmente.

Unir los archivos. Proceso que une los 88 archivos CSV etiquetados para formar un dataset[4] global. Una vez se tiene un solo archivo CSV con toda la información, se procedió a guardarlo como un archivo de texto delimitado por tabulaciones, este último paso es opcional, sin embargo, es más fácil trabajar con archivos de texto que con archivos CSV. Para este proceso se utilizó la herramienta de Python.

Limpieza de datos. Proceso en el que se debe limpiar el dataset, para ello se utilizó las librerías “Pandas”, “String”, “Re” de Python. A continuación, se detallan las tareas que se realizaron:

- **Remover espacios en blanco:** Aquí se utilizó (string.whitespace)
- **Remover caracteres especiales:** En este apartado se optó por el uso de expresiones regulares que removieron los siguientes caracteres ('!"#$&*+:;<=>?@[\\]^_`{|}').
- **Convertir a minúscula:** Aquí se utilizó el método (lower())

Carga. Una vez se obtiene el dataset totalmente limpio, se procede al tratamiento de este para que cumpla con el formato adecuado previo el entrenamiento del modelo machine learning.

Transformación a formato spaCy. Para entrenar modelos NER con la librería de spaCy, se necesita un formato especifico que puede ser transformado fácilmente desde el

[4] dataset: Conjunto de datos

dataset que se tiene actualmente. El formato al que deben ajustarse los datos es una lista de tuplas que contienen los siguientes cuatro atributos ***(Pardeshi, 2020)*** y que se los muestra en la figura 22:

i. Texto.

ii. Posición inicial de la palabra que hace referencia la entidad.

iii. Posición final de la palabra que hace referencia la entidad.

iv. La etiqueta que corresponde a la entidad.

Figura 22

Ejemplo formato spaCy

```
[("characterization of silver nanoparticles by green
synthesis method using pedalium murex leaf extract and
their antibacterial activity abstract in",
  {'entities': [(20, 26, 'B-PSN_TIPNPSOBT'),
    (73, 81, 'B-AIA_ESPECIE'),
    (82, 87, 'I-AIA_ESPECIE'),
    (88, 92, 'B-AIA_PARTE'),
    (93, 100, 'B-PSN_AR'),
    (111, 124, 'B-AP_ACTCAP'),
    (125, 133, 'I-AP_ACTCAP')]
  }
 )
]
```

Nota. Este gráfico muestra un ejemplo de cómo se tienen que transformar los datos a un formato que aceptan los modelos de spaCy.

Dividir el dataset. Los algoritmos machine learning aprenden de los datos que se les provea para el entrenamiento, a partir de ellos estos algoritmos intentan encontrar patrones para así predecir un resultado. Para comprobar la validez de un modelo se necesita de igual

manera más datos, por lo tanto, se tiene que dividir los datos en conjuntos de entrenamiento y conjuntos de prueba (Recuero, 2022).

El conjunto de datos de entrenamiento es el que se utiliza para entrenar al modelo, la calidad del modelo depende de la calidad de los datos, por esta razón, la etapa de ETL es la que consumió más tiempo.

El conjunto de datos de prueba es el que se utiliza para comprobar si el modelo funciona de manera correcta, es decir, si el modelo es capaz de predecir correctamente a la tarea que se le fue entrenado.

Normalmente se suele dividir el 80% de los datos para el conjunto de entrenamiento y 20% para el conjunto de prueba, además, se toman muestras aleatorias, es decir, se mezclan todos los datos y luego se los dividen (Sets de Entrenamiento, Test y Validación, 2020).

Por lo mencionado anteriormente, primero se utilizó la librería “random” de Python para aleatorizar los datos y después se los dividieron. Sin embargo, para esta última tarea, debido a la reducida cantidad de datos que se poseía, se dividió el 90% de los datos para el conjunto de entrenamiento y un 10% para el conjunto de prueba.

Después, se guardaron estos conjuntos de datos en archivos de extensión “pickle”, la razón de esto es debido a que con ellos se pueden guardar objetos de Python sin la necesidad de realizar algún tipo de conversión. Finalmente, se los transforman a archivos con extensión (.spacy).

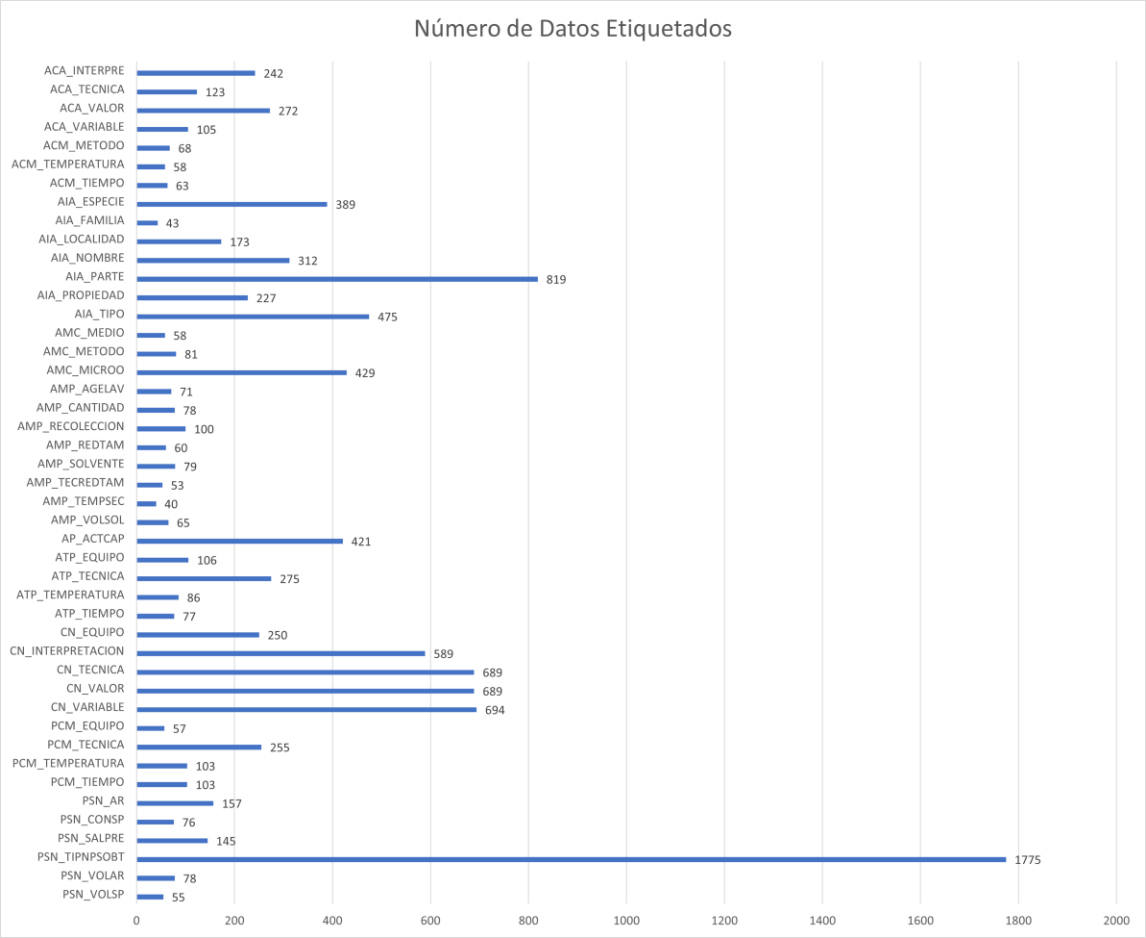
Entrenamiento del modelo

La fase de entrenamiento es aquella donde, a través de los datos recolectados y transformados a un dataset compacto, se los utiliza para el entrenamiento y posterior validación del modelo machine learning.

Recopilando la información, se logró obtener 88 artículos científicos, por lo tanto, el dataset cuenta con 3399 oraciones, 289116 palabras y de las cuales 11163 datos están etiquetados. A continuación, se detalla el número de datos etiquetados de cada entidad.

Figura 23

Etiquetas dentro del dataset



Nota. Este gráfico muestra el número de etiquetas de cada entidad.

Elección del modelo. Como se mencionó anteriormente, para realizar el modelo, se utilizó la librería spaCy de Python ya que tiene varios algoritmos relacionados al Procesamiento del Lenguaje Natural (NLP).

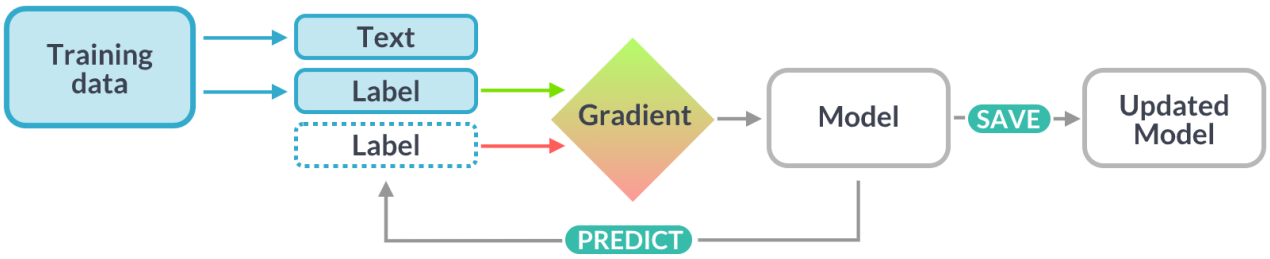
Para el entrenamiento de los modelos, spaCy utiliza leyes estadísticas que dependen de los datos seleccionados para el entrenamiento. Cada “decisión” que toman estos modelos estadísticos, por ejemplo, saber que etiqueta asignar a cada palabra o si una palabra representa una entidad, es una predicción basada en los valores de ponderación actuales del modelo (Explotion AI, 2022). La predicción se basa en los ejemplos que el modelo ha visto durante el entrenamiento.

El entrenamiento es un proceso iterativo en el que las predicciones del modelo se comparan con las anotaciones de referencia para estimar el gradiente de perdida. El gradiente de perdida se utiliza para calcular el gradiente de los pesos. Los gradientes indican como se deben cambiar los valores de pesos para mejorar la predicción (Explotion AI, 2022). A continuación, se mostrará la arquitectura (ver figura 24) y componentes de entrenamiento de un modelo spaCy:

- **Datos de entrenamiento:** Ejemplos y anotaciones.
- **Texto:** El texto de entrada para el que el modelo debe predecir una etiqueta.
- **Etiqueta:** La etiqueta que el modelo debe predecir.
- **Gradiente:** La dirección y la taza de cambio de un valor numérico. Minimizar el gradiente de los pesos genera predicciones más cercanas a las etiquetas de referencia en los datos de entrenamiento.

Figura 24

Arquitectura del entrenamiento



Nota. Este gráfico muestra la arquitectura de entrenamiento de un modelo spaCy. Adaptado de *Training Pipelines & Models*, por Explotion AI, 2022 (https://spacy.io/usage/training#basics). Derechos de autor 2022 por Explotion AI.

La librería de spaCy cuenta con varios módulos, en este caso se utilizó el módulo NER, este se encarga de identificar la ubicación de una frase dentro del texto y determinar el tipo de entidad al que hace referencia.

Además, spaCy incorpora muchos modelos optimizados para diferentes idiomas, en este caso se personalizó un modelo NER utilizando el modelo ‘en’ que hace referencia al idioma inglés.

Para entrenar un modelo NER personalizado sin un modelo previamente entrenado, normalmente se necesita tener alrededor de 2000 a 1 millón de ejemplos tanto para el entrenamiento como para la evaluación. Entonces, debido a esto último y tomando en cuenta la cantidad de datos que se posee, se realizó el entrenamiento de dos modelos NER personalizados con las etiquetas presentadas previamente en la tabla 8, esto con la finalidad de combinar lo mejor de ambas tecnologías para obtener mejores resultados.

El primer modelo NER proporcionado por spaCy consta de una sofisticada técnica de incrustación de palabras (word embedding) que utiliza características de subpalabras e incrustaciones “Bloom”, una red neuronal convolucional profunda con conexiones residuales y

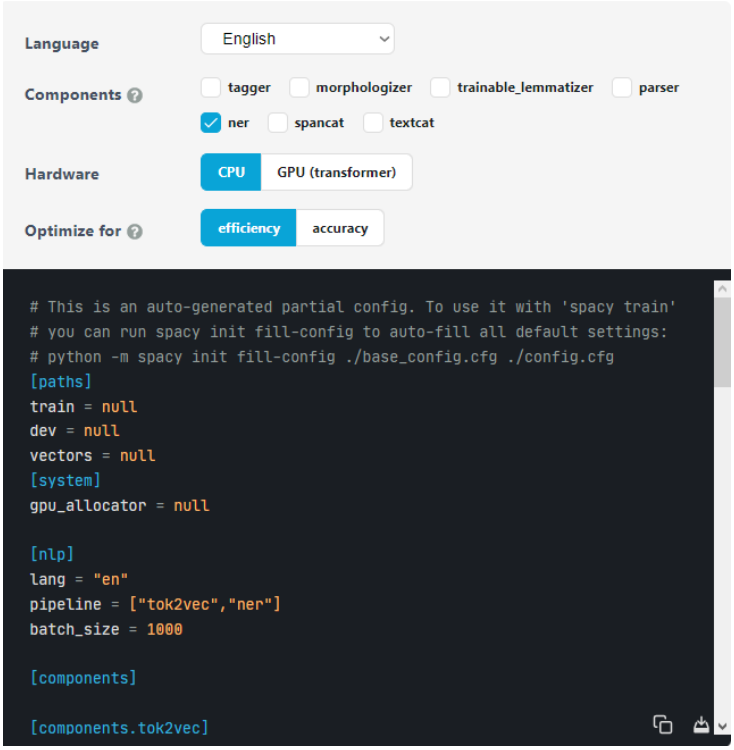
un enfoque novedoso basado en transiciones para el análisis de entidades nombradas. El sistema está diseñado para ofrecer un equilibrio entre eficiencia, precisión y adaptabilidad (Honnibal, 2017).

El segundo modelo NER proporcionado por spaCy consiste en una arquitectura de transformadores de última generación, en este caso se utilizó un transformador pre entrenado denominado RoBERTa que es una versión mejorada de BERT.

Entrenamiento del módulo NER. Los dos modelos que se utilizaron fueron entrenados de la misma manera con la diferencia de la elección del tipo de modelo, a continuación, se mostrará el procedimiento necesario para entrenar el módulo NER de spaCy:

Figura 25

Configuración del primer modelo



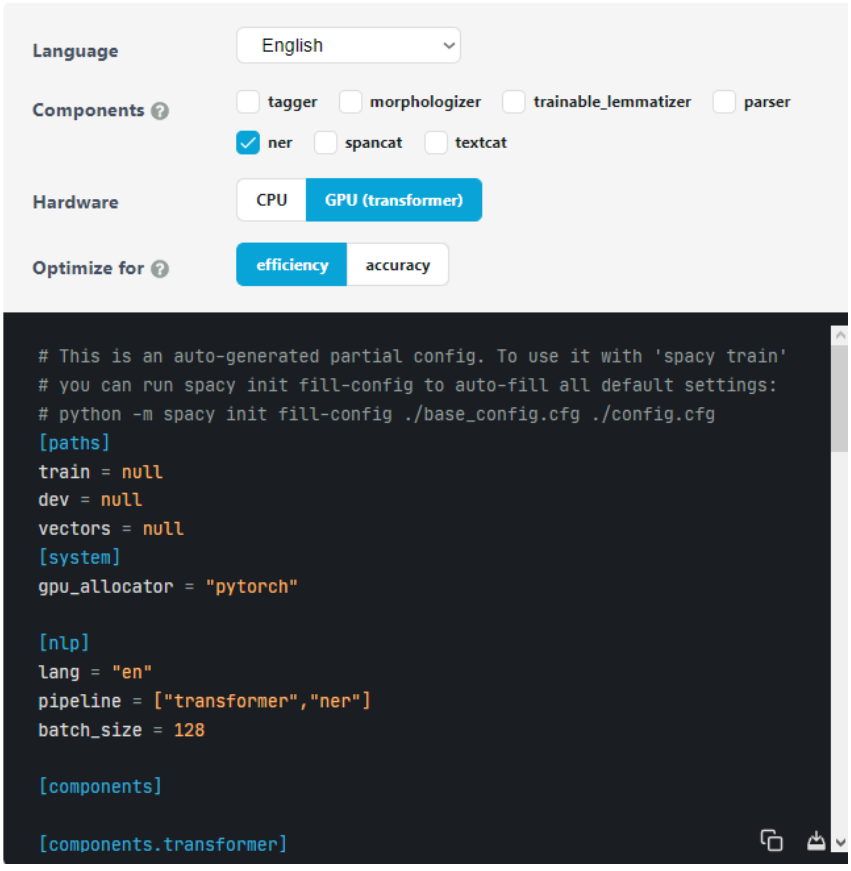
Nota. Este gráfico muestra la configuración del primer modelo spaCy. Adaptado de *Training Pipelines & Models*, por spaCy, 2022 (https://spacy.io). Derechos de autor 2022 por spaCy.

Primero se descargó los archivos de configuración de los modelos desde la página oficial de spaCy. Este paso es muy importante ya que, a través de lo que se seleccione se configurara el modelo.

Para el primer modelo se seleccionó el idioma inglés, con el componente de NER y optimizado para la eficiencia como se puede observar en la figura 25. Para el segundo modelo se seleccionó lo mismo con la diferencia de que se debe seleccionar la opción de los transformadores como se puede observar en la figura 26.

Figura 26

Configuración del segundo modelo



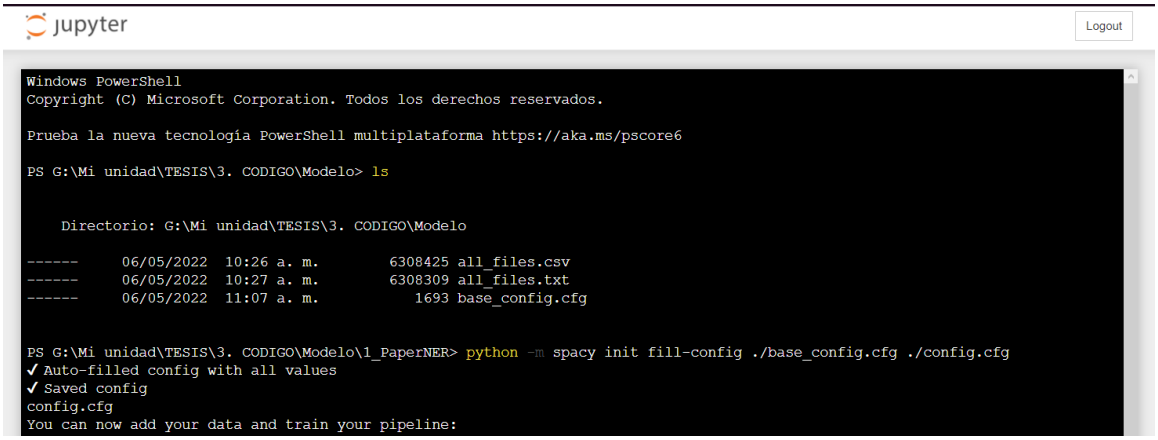
Nota. Este gráfico muestra la configuración del segundo modelo spaCy. Adaptado de Training Pipelines & Models, por spaCy, 2022 (https://spacy.io). Derechos de autor 2022 por spaCy.

Estos archivos son muy importantes ya que establecen un punto de partida con los parámetros e hiperparámetros recomendados de cada módulo, en este caso el módulo NER.

Una vez descargado estos archivos de configuración, en el entorno de trabajo se debe ejecutar el siguiente comando (python -m spacy init fill-config ./base_config.cfg ./config.cfg) para cargar el archivo de configuración del modelo seleccionado (ver figura 27).

Figura 27

Carga del archivo de configuración del modelo



Nota. Este gráfico muestra cómo se debe cargar el archivo de configuración de los modelos spaCy.

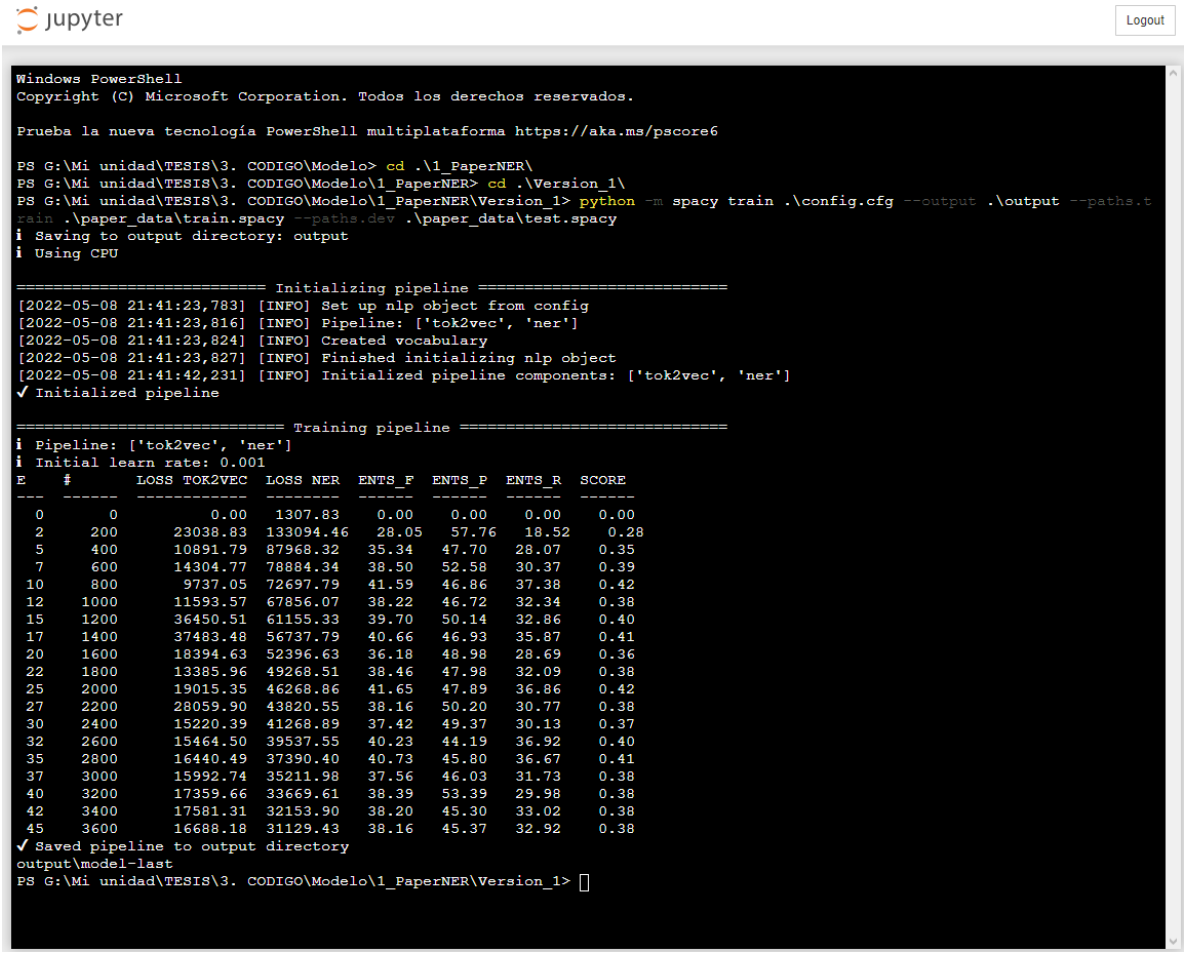
Finalmente, para el entrenamiento del modelo se tendrá que crear una carpeta denominada “output” que es donde se almacenaran los modelos luego del entrenamiento. Esta carpera almacena dos carpetas, la primera “model-best” hace referencia al modelo con mejor rendimiento durante el entrenamiento mientras que la segunda “model-last” hace referencia al último modelo en ser entrenado. En este caso, para las pruebas se utilizó el modelo con mejor rendimiento durante el entrenamiento, es decir, el denominado “model-best”.

Para comenzar con el entrenamiento se debe ejecutar el comando que especifica la ruta donde se encuentra el archivo de configuración y finalmente la ruta donde se encuentra el

dataset de entrenamiento y pruebas (python -m spacy train .\config.cfg --output .\output --paths.train .\paper_data\train.spacy --paths.dev .\paper_data\test.spacy). Este procedimiento se realizó para los dos modelos NER seleccionados y se obtuvieron resultados similares a los presentados en la figura 28.

Figura 28

Ejemplo de los resultados del entrenamiento



Nota. Este gráfico muestra los resultados del entrenamiento de los modelos de spaCy.

Capítulo IV

En esta sección se presentan los resultados obtenidos con los modelos machine learning que fueron entrenados y además la validación del modelo por parte del personal del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE.

Evaluación del Rendimiento del modelo

Una vez entrenados los modelos machine learning que se detallaron en la sección anterior, el siguiente paso y también importante, es realizar la evaluación del rendimiento de los modelos.

Cada vez que se construye un modelo machine learning, es necesario contar con métricas que evalúen el rendimiento del modelo, determinar si un modelo tiene éxito en la tarea al que fue entrenado depende de dos factores (Pykes, 2021):

- Si la métrica de evaluación que se seleccionó es la correcta para el problema.
- Si se está siguiendo el proceso de evaluación correcto.

Matriz de confusión

En el campo del Procesamiento del Lenguaje Natural (NLP) y específicamente para tareas NER se suele utilizar las matrices de confusión como una herramienta de visualización de rendimiento de los modelos machine learning.

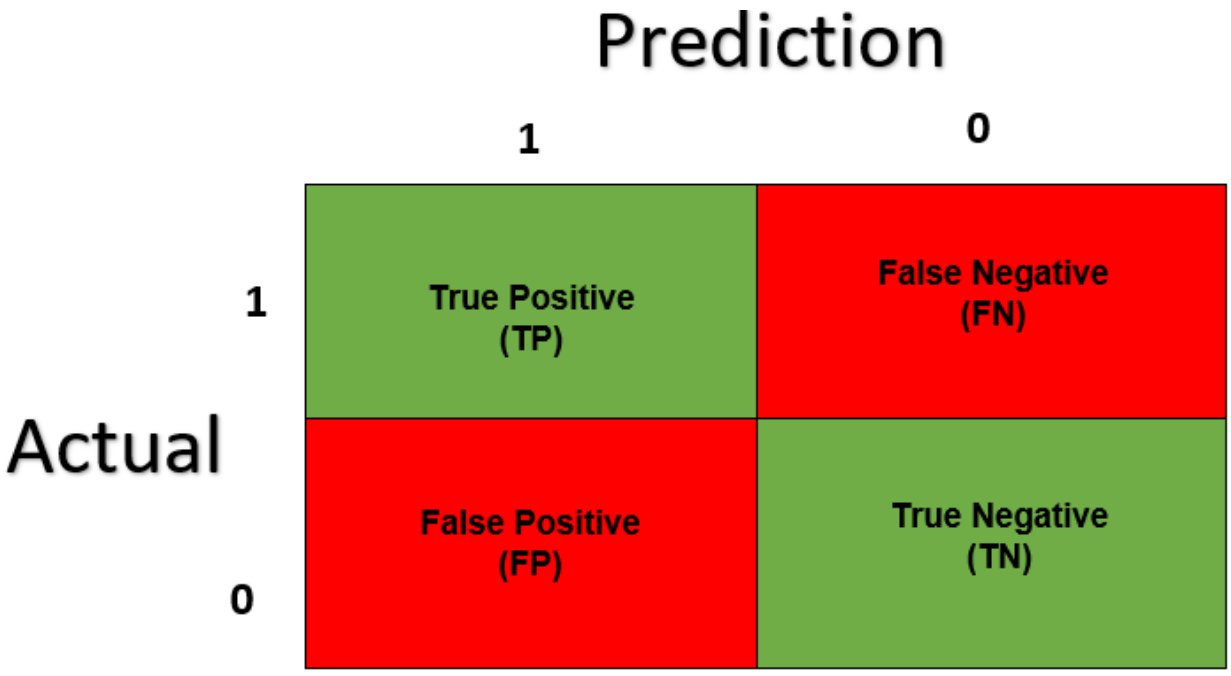
Una matriz de confusión es una representación matricial de los resultados de las predicciones de cualquier prueba binaria que se utiliza a menudo para describir el rendimiento de modelos clasificadores sobre un conjunto de datos de prueba cuyos valores reales se conocen (Singh, 2020).

Además, la matriz de confusión está compuesta por cuatro elementos muy importantes que posteriormente servirán como variables para las métricas de evaluación. A continuación, se muestra los componentes y sus respectivos significados:

- **Verdadero Positivo (TP):** El modelo predijo positivo y la etiqueta fue realmente positiva.
- **Verdadero Negativo (TN):** El modelo predijo negativo y la etiqueta fue realmente negativa.
- **Falso Positivo (FP):** El modelo predijo positivo y la etiqueta fue negativa.
- **Falso Negativo (FN):** El modelo predijo negativo y la etiqueta fue positiva.

Figura 29

Componentes de una matriz de confusión



Nota. Este gráfico demuestra los componentes de una matriz de confusión. Adaptado de *Confusion Matrix "Un-confused",* por C. Pykes, 2020

(https://towardsdatascience.com/confusion-matrix-un-confused-1ba98dee0d7f). Derechos de autor 2020 por Towards Data Science.

La matriz de confusión crea un resumen de los resultados para el modelo de predicción, en este caso para los modelos NER, y además sirve para realizar análisis sobre estos resultados.

Para este caso se obtuvieron dos matrices de confusión ya que se entrenaron dos modelos NER, el primero con Redes Neuronales Convolucionales (CNN) y el segundo con RoBERTa.

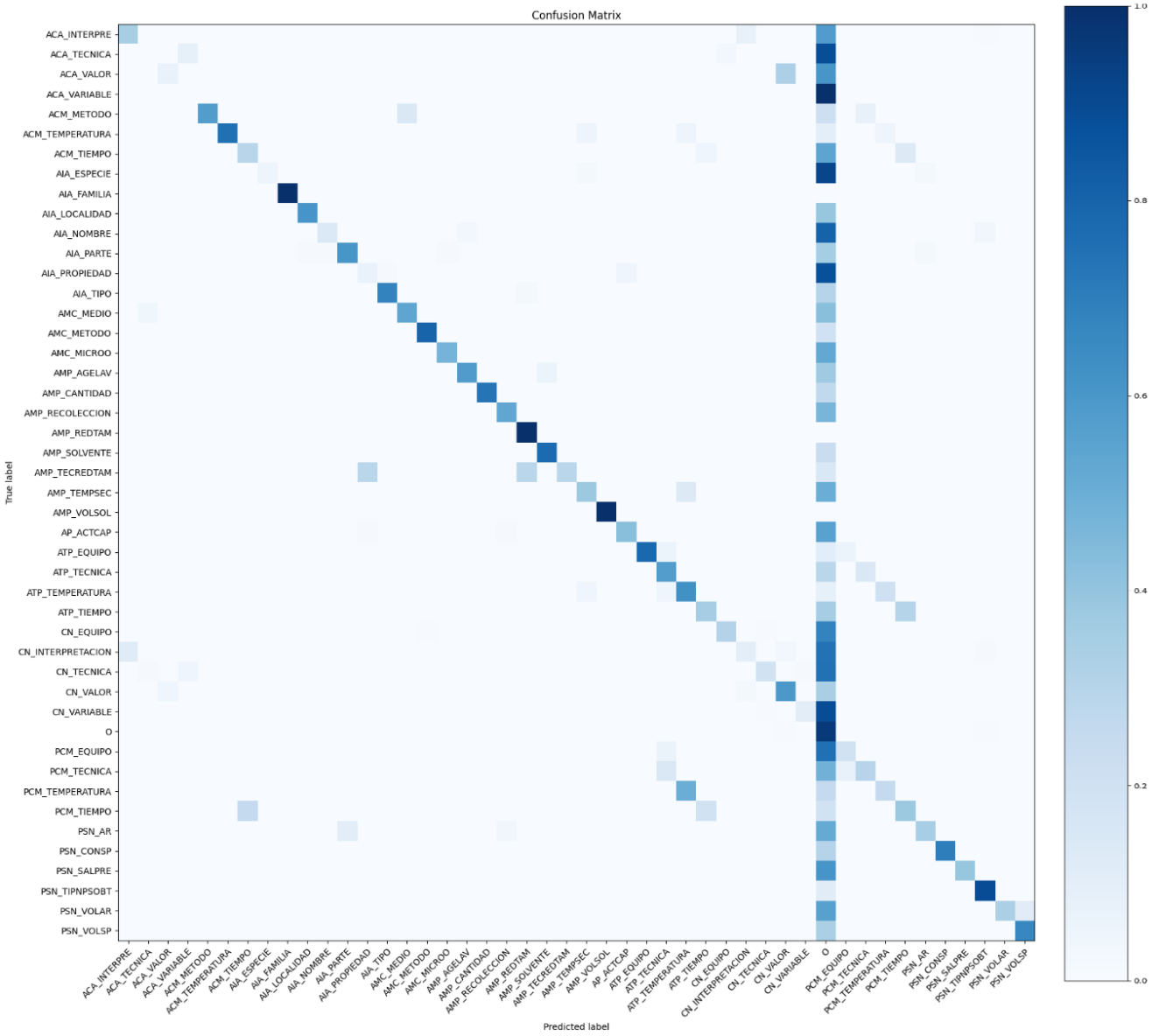
Las matrices que se elaboraron están compuestas por valores del 0 al 1, donde los cuadros con colores azules más oscuros representan valores cercanos al 1 mientras que los cuadros de color azul más claros hacen referencia a valores cercanos al 0. Los elementos diagonales representan el número de etiquetas predichas correctamente por el modelo mientras que los datos que se encuentran fuera de la diagonal son predicciones incorrectas.

Como se puede observar en la figura 30, existen etiquetas que son predichas correctamente por el modelo, sin embargo, también se puede observar que realiza predicciones incorrectas. Lo mismo ocurre en la figura 31, pero con el detalle de que en este se predicen algunas etiquetas que en la imagen anterior (modelo con CNN) no pudo predecir.

Estas matrices de confusión permiten observar de manera rápida como están funcionando los modelos machine learning, sin embargo, más adelante se detalla cada uno de estos resultados gracias a la función “evaluate” de la librería de spaCy de Python. Al ejecutar este comando (python -m spacy evaluate output/model-best paper_data/test.spacy --output output/metrics.json) se obtiene una evaluación detallada de cada etiqueta.

Figura 30

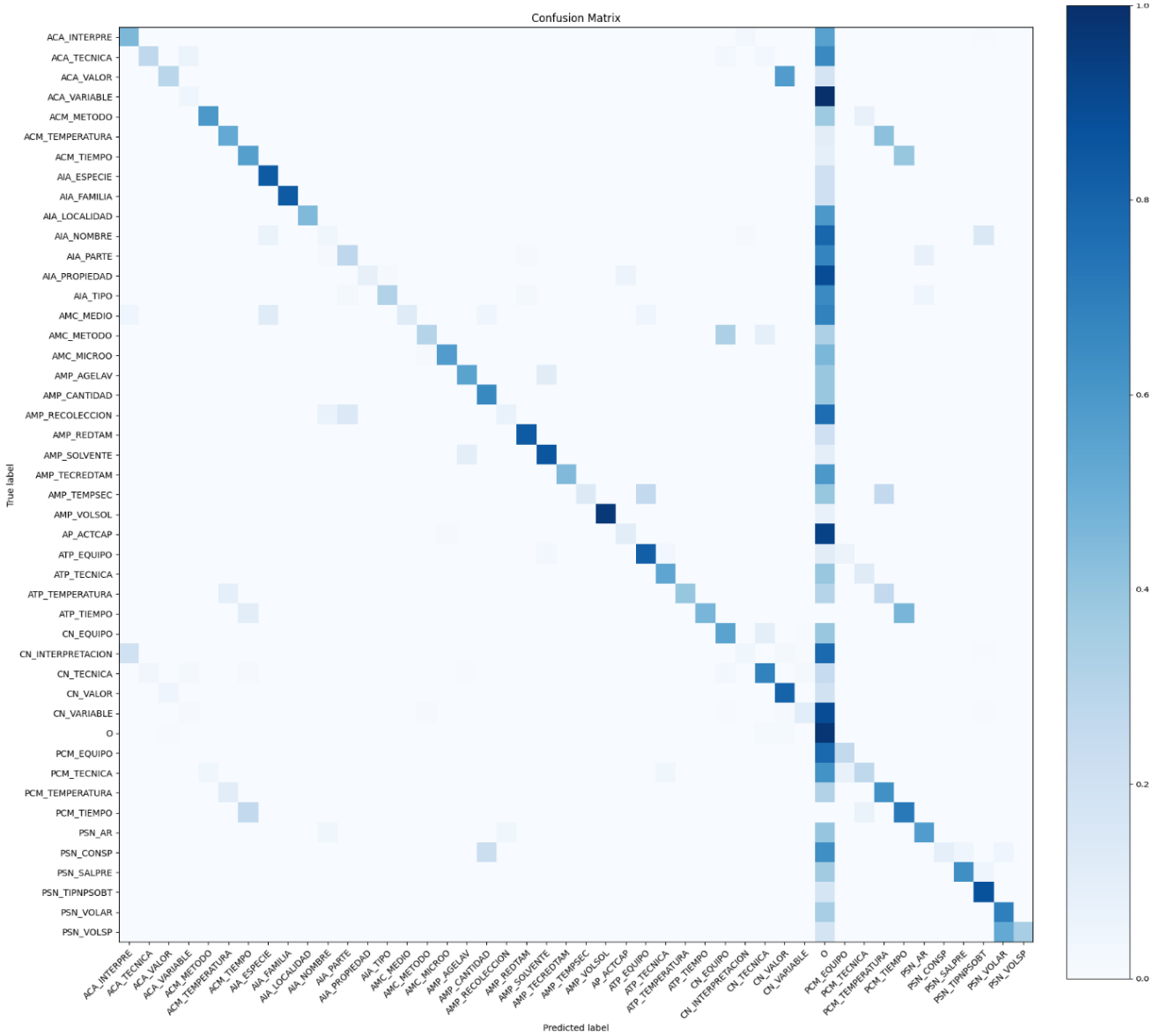
Matriz de confusión del primer modelo con CNN



Nota. Este gráfico muestra la matriz del primer modelo.

Figura 31

Matriz de confusión del segundo modelo con RoBERTa



Nota. Este gráfico muestra la matriz del segundo modelo.

Métricas de evaluación

Los modelos NLP se suelen evaluar con respecto a su rendimiento sobre un conjunto de prueba, por ello se utilizaron métricas de evaluación intrínsecas que son las siguientes:

- **Precisión:** Mide cuan exactas son las predicciones del modelo. Esta métrica informa el número de etiquetas que en realidad están etiquetadas correctamente.

$$P = \frac{TP}{TP + FP}$$

- **Recall:** Mide que tan bien el modelo puede recordar la clase positiva, es decir, el número de etiquetas correctas que el modelo identifico como correctas.

$$R = \frac{TP}{TP + FN}$$

- **Puntuación F1:** La precisión y recall son métricas complementarias que tienen una relación inversa. Si ambos nos interesan se usa la puntuación F1 para combinar la precisión y el recall en una sola métrica y así obtener una estimación de la calidad general del modelo.

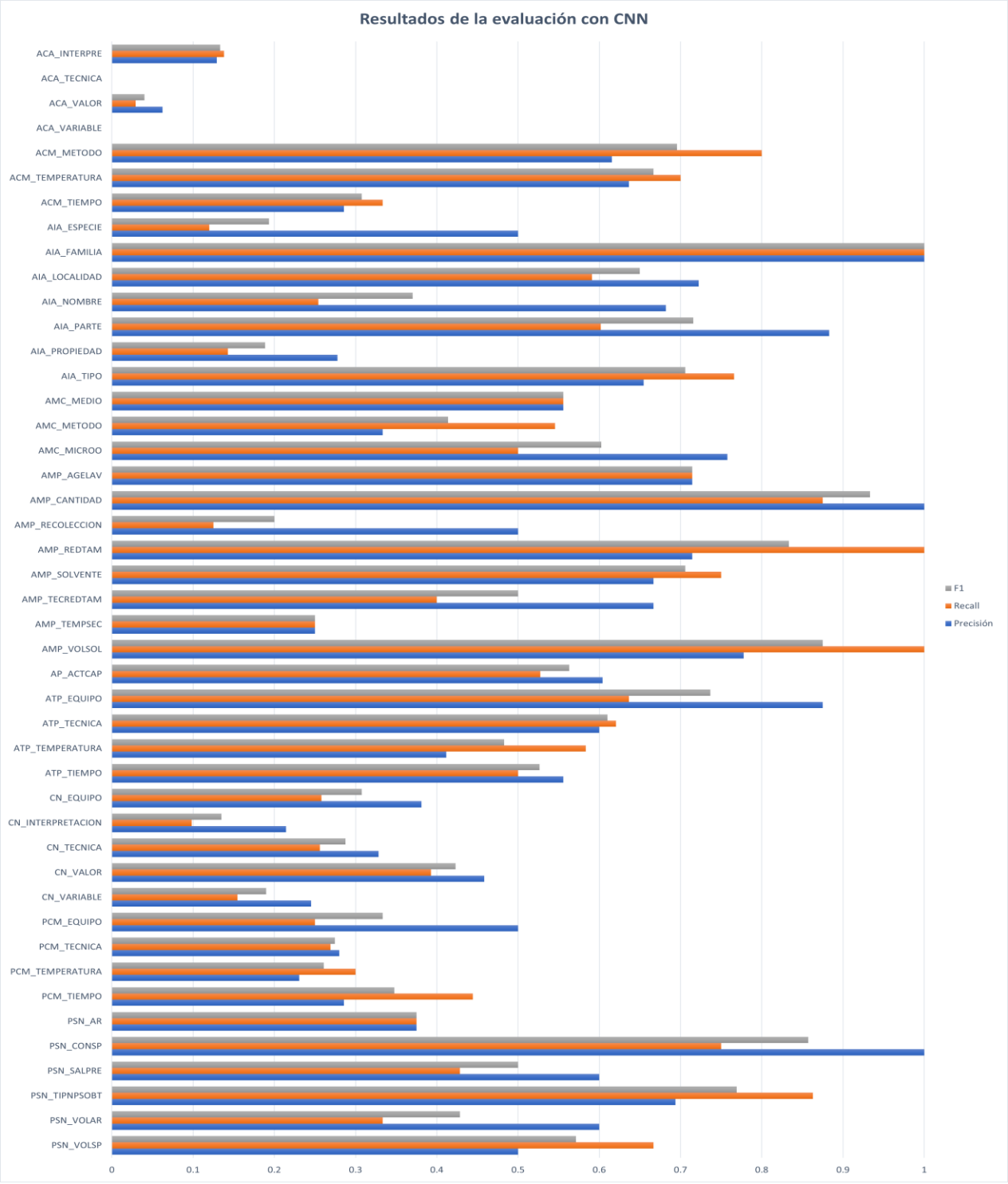
$$F1 = 2\frac{P \times R}{P + R}$$

Como se mencionó anteriormente, spaCy cuenta con una función denominada “evaluate” que nos proporciona una evaluación de cada una de las etiquetas, en dicha evaluación se detallan las métricas mencionadas (precisión, recall y F1).

En la figura 32 se puede observar los resultados obtenidos de cada una de las etiquetas con el modelo NER con CNN, donde las franjas de color plomo, naranja y azul representan al puntaje F1, recall y precisión respectivamente. De igual manera, en la figura 33 se observa los resultados obtenidos con el modelo NER con RoBERTa.

Figura 32

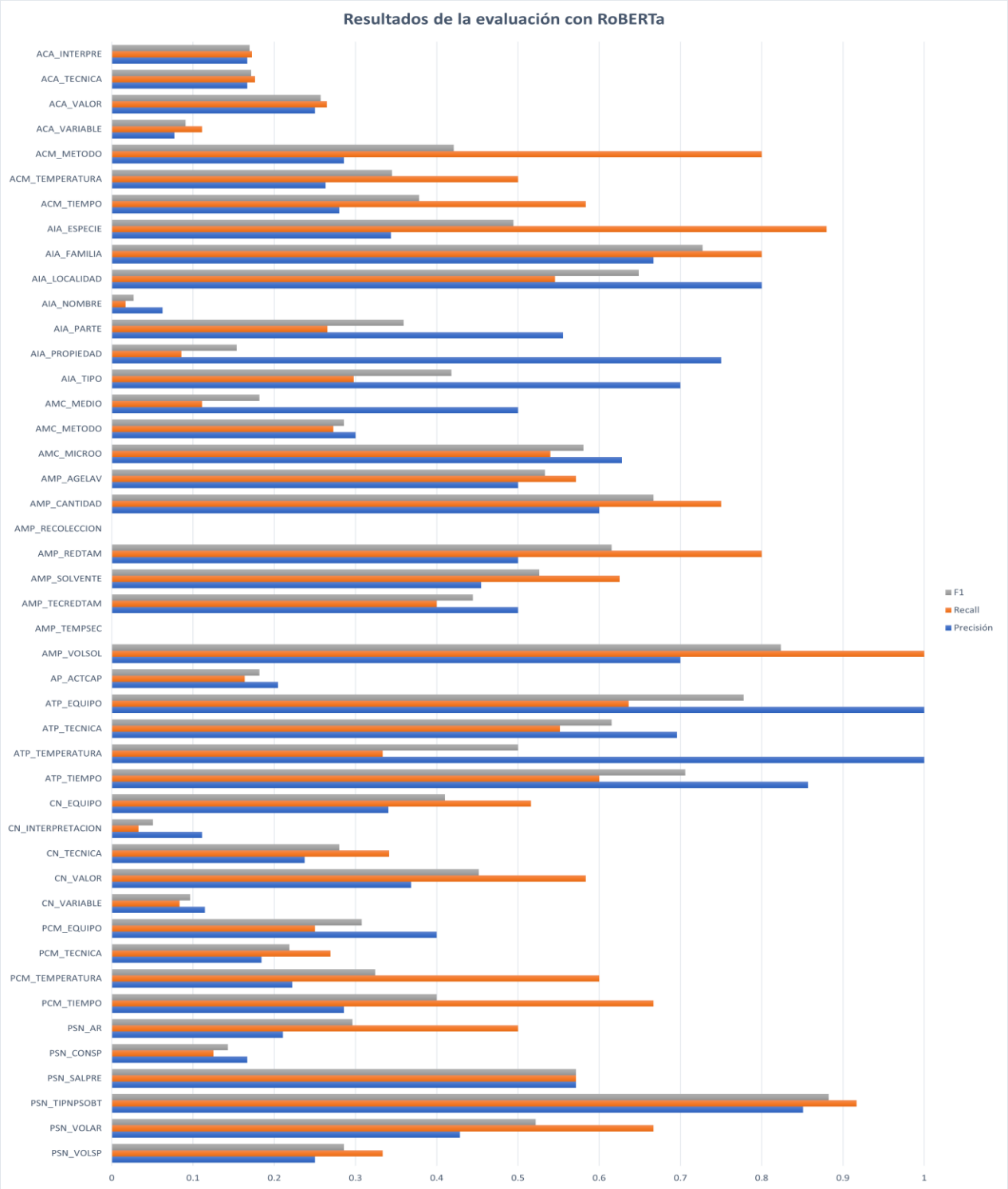
Resultados obtenidos con el modelo NER con CNN



Nota. Este gráfico muestra los resultados del primer modelo.

Figura 33

Resultados obtenidos con el modelo NER con RoBERTa



Nota. Este gráfico muestra los resultados del segundo modelo.

Análisis de resultados

A continuación, se detallan los resultados observados en las gráficas anteriores. Se analizó la precisión con la que ambos modelos lograron predecir correctamente, específicamente, en la tabla 9 se resaltan las etiquetas que lograron una precisión superior al 50% junto con su porcentaje de anotaciones dentro del dataset.

Tabla 9

Análisis de la precisión de ambos modelos

N	Modelo CNN	Modelo RoBERTa	Porcentaje de anotaciones
1	AIA_TIPO	AIA_TIPO	4.26%
2	AIA_NOMBRE	AIA_NOMBRE	2.79%
3	AIA_ESPECIE	AIA_ESPECIE	3.48%
4	AIA_FAMILIA	AIA_FAMILIA	0.39%
5	AIA_PARTE	AIA_PARTE	7.34%
6	AIA_LOCALIDAD	AIA_LOCALIDAD	1.55%
7	AIA_PROPIEDAD	AIA_PROPIEDAD	2.03%
8	AMP_RECOLECCION	AMP_RECOLECCION	0.90%
9	AMP_AGELAV	AMP_AGELAV	0.64%
10	AMP_TEMPSEC	AMP_TEMPSEC	0.36%

N	Modelo CNN	Modelo RoBERTa	Porcentaje de anotaciones
11	AMP_TECREDTAM	AMP_TECREDTAM	0.47%
12	AMP_REDTAM	AMP_REDTAM	0.54%
13	AMP_CANTIDAD	AMP_CANTIDAD	0.70%
14	AMP_SOLVENTE	AMP_SOLVENTE	0.71%
15	AMP_VOLSOL	AMP_VOLSOL	0.58%
16	ATP_TECNICA	ATP_TECNICA	2.46%
17	ATP_TIEMPO	ATP_TIEMPO	0.69%
18	ATP_TEMPERATURA	ATP_TEMPERATURA	0.77%
19	ATP_EQUIPO	ATP_EQUIPO	0.95%
20	ACA_TECNICA	ACA_TECNICA	1.10%
21	ACA_VARIABLE	ACA_VARIABLE	0.94%
22	ACA_VALOR	ACA_VALOR	2.44%
23	ACA_INTERPRE	ACA_INTERPRE	2.17%
24	PSN_AR	PSN_AR	1.41%
25	PSN_VOLAR	PSN_VOLAR	0.70%

N	Modelo CNN	Modelo RoBERTa	Porcentaje de anotaciones
26	PSN_SALPRE	PSN_SALPRE	1.30%
27	PSN_VOLSP	PSN_VOLSP	0.49%
28	PSN_CONSP	PSN_CONSP	0.68%
29	PCM_TECNICA	PCM_TECNICA	2.28%
30	PCM_EQUIPO	PCM_EQUIPO	0.51%
31	PCM_TEMPERATURA	PCM_TEMPERATURA	0.92%
32	PCM_TIEMPO	PCM_TIEMPO	0.92%
33	PSN_TIPNPSOBT	PSN_TIPNPSOBT	15.90%
34	CN_TECNICA	CN_TECNICA	6.17%
35	CN_VARIABLE	CN_VARIABLE	6.22%
36	CN_VALOR	CN_VALOR	6.17%
37	CN_INTERPRETACION	CN_INTERPRETACION	5.28%
38	CN_EQUIPO	CN_EQUIPO	2.24%
39	AP_ACTCAP	AP_ACTCAP	3.77%
40	AMC_METODO	AMC_METODO	0.73%

N	Modelo CNN	Modelo RoBERTa	Porcentaje de anotaciones
41	AMC_MICROO	AMC_MICROO	3.84%
42	AMC_MEDIO	AMC_MEDIO	0.52%
43	ACM_METODO	ACM_METODO	0.61%
44	ACM_TEMPERATURA	ACM_TEMPERATURA	0.52%
45	ACM_TIEMPO	ACM_TIEMPO	0.56%

Nota. Esta tabla muestra las etiquetas que obtuvieron un puntaje superior al 50% en su precisión de ambos modelos junto con su porcentaje de anotaciones dentro del dataset.

Como se puede observar, en el modelo NER con CNN, 21 etiquetas logran una precisión superior al 50% y, por otra parte, en el modelo NER con RoBERTa, 18 etiquetas logran una precisión superior al 50%.

Cabe resaltar que existen etiquetas que fueron predichas correctamente por el modelo CNN que en el modelo con RoBERTa no logró tal precisión y así viceversa, es por esta razón que al momento de presentar una extracción final se unió los resultados de ambos modelos para de esta manera enriquecer con más información.

No obstante, también existen etiquetas que no lograron cumplir con la predicción esperada, por ello a continuación, se detallan las razones del porque las etiquetas restantes decaen en sus métricas de evaluación:

- **Distribución de datos desequilibrada:** Como se pudo observar en la figura 23, los datos se encuentran desequilibrados, es decir, no todas las etiquetas cuentan con la misma proporción de anotaciones. Por tal motivo por ejemplo las

etiquetas AIA_TIPO, AIA_NOMBRE, AIA_PARTE, AIA_LOCALIDAD, AIA_PROPIEDAD, ATP_TECNICA, PSN_SALPRE, PSN_TIPNPSOBT, AP_ACTCAP y AMC_MICROO tienen métricas de evaluación altas.

- **Ambigüedad:** Existe ambigüedad entre algunas etiquetas que extraen información parecida como por ejemplo la etiqueta AMP_SOLVENTE y AMP_AGELAV; ATP_TECNICA y PCM_TECNICA, entre otras.
- **Valores numéricos:** Etiquetas referentes a valores numéricos como por ejemplo la etiquetas AMP_TEMPSEC, ACA_VALOR, PSN_VOLAR, PSN_VOLSP, PSN_CONSP, entre otras.
- **Frases largas:** Existen etiquetas que deben extraer interpretaciones y frases largas. Generalmente los modelos NER funcionan mejor para detectar frases relativamente cortas (Prodigy, 2022). Las etiquetas CN_INTERPRETACION, ACA_INTERPRE, entre otras, forman parte de este problema.
- **Dispersión de datos en el texto:** Existen etiquetas que deben extraer datos de diferentes partes del texto para formar una sola, por ello, las etiquetas ACA_TECNICA, ACA_VARIABLE, ACA_VALOR, CN_TECNICA, CN_VARIABLE, CN_VALOR, CN_EQUIPO, entre otras, no tienen una buena puntuación en sus métricas de evaluación.

Finalmente, otro factor que influyó es el número de entidades dentro del entrenamiento puesto que, al tener 45 entidades, a mayor número de entidades contenidas en los datos del entrenamiento a igualdad de textos de entrenamiento, y con esto mayor variabilidad es introducida, es decir, al contar con un dataset relativamente reducido el rendimiento de los modelos NER tiende a errar. Por esta razón se suele recomendar entrenar modelos NER con la

menor cantidad posible de entidades para asegurar mejores resultados en las métricas de evaluación.

Validación del modelo

Para validar los resultados obtenidos con los modelos machine learning, se realizó una reunión con el personal del área de Biotecnología de la Universidad de las Fuerzas Armadas – ESPE., quienes además fueron los que participaron al inicio de este proyecto.

Entonces, con la finalidad de comprobar los resultados obtenidos por los modelos, se realizó una comparación entre los datos obtenidos con los modelos y la base de datos en Excel que contenía la extracción manual de los datos. Para esta tarea se utilizaron 4 artículos científicos que no formaban parte del dataset, la razón de esto es porque si se usaba algún artículo científico que fue utilizado para las fases de entrenamiento y prueba el modelo ya conocería el resultado esperado lo cual sesgaría los resultados. De esta manera se obtuvieron los resultados que se visualizan en la tabla 10, donde se detallan los siguientes términos:

- **Correcto:** Los modelos y la base de datos coinciden.
- **Neutral:** Los modelos y la base de datos coinciden parcialmente, es decir, los modelos no lograron extraer por completo la información.
- **Incorrecto** Los modelos y la base de datos no coinciden.

Tabla 10

Comparación de resultados

N	Etiquetas	X88	X167	X120	X9
1	AIA_TIPO	Correcto	Correcto	Correcto	Correcto

N	Etiquetas	X88	X167	X120	X9
2	AIA_NOMBRE	Correcto	Correcto	Correcto	Correcto
3	AIA_ESPECIE	Incorrecto	Correcto	Incorrecto	Correcto
4	AIA_FAMILIA	Correcto	Incorrecto	Incorrecto	Correcto
5	AIA_PARTE	Incorrecto	Incorrecto	Correcto	Correcto
6	AIA_LOCALIDAD	Correcto	Correcto	Incorrecto	Correcto
7	AIA_PROPIEDAD	Incorrecto	Correcto	Incorrecto	Correcto
8	AMP_RECOLECCION	Correcto	Correcto	Incorrecto	Correcto
9	AMP_AGELAV	Incorrecto	Correcto	Incorrecto	Incorrecto
10	AMP_TEMPSEC	Correcto	Incorrecto	Incorrecto	Correcto
11	AMP_TECREDTAM	Correcto	Correcto	Neutral	Correcto
12	AMP_REDTAM	Incorrecto	Correcto	Correcto	Correcto
13	AMP_CANTIDAD	Correcto	Correcto	Incorrecto	Correcto
14	AMP_SOLVENTE	Correcto	Correcto	Incorrecto	Correcto
15	AMP_VOLSOL	Incorrecto	Correcto	Incorrecto	Correcto
16	ATP_TECNICA	Correcto	Correcto	Neutral	Correcto
17	ATP_TIEMPO	Correcto	Incorrecto	Incorrecto	Correcto

N	Etiquetas	X88	X167	X120	X9
18	ATP_TEMPERATURA	Correcto	Incorrecto	Incorrecto	Correcto
19	ATP_EQUIPO	Correcto	Correcto	Incorrecto	Correcto
20	ACA_TECNICA	Incorrecto	Neutral	Incorrecto	Correcto
21	ACA_VARIABLE	Incorrecto	Incorrecto	Neutral	Incorrecto
22	ACA_VALOR	Incorrecto	Neutral	Incorrecto	Neutral
23	ACA_INTERPRE	Correcto	Neutral	Neutral	Neutral
24	PSN_AR	Correcto	Correcto	Incorrecto	Correcto
25	PSN_VOLAR	Correcto	Correcto	Incorrecto	Correcto
26	PSN_SALPRE	Correcto	Correcto	Incorrecto	Correcto
27	PSN_VOLSP	Correcto	Correcto	Incorrecto	Correcto
28	PSN_CONSP	Correcto	Correcto	Incorrecto	Correcto
29	PCM_TECNICA	Correcto	Incorrecto	Neutral	Incorrecto
30	PCM_EQUIPO	Correcto	Correcto	Incorrecto	Correcto
31	PCM_TEMPERATURA	Correcto	Incorrecto	Incorrecto	Incorrecto
32	PCM_TIEMPO	Correcto	Incorrecto	Incorrecto	Incorrecto
33	PSN_TIPNPSOBT	Correcto	Correcto	Incorrecto	Correcto

N	Etiquetas	X88	X167	X120	X9
34	CN_TECNICA	Correcto	Neutral	Incorrecto	Neutral
35	CN_VARIABLE	Neutral	Neutral	Neutral	Neutral
36	CN_VALOR	Neutral	Neutral	Neutral	Neutral
37	CN_INTERPRETACION	Neutral	Neutral	Neutral	Neutral
38	CN_EQUIPO	Incorrecto	Neutral	Incorrecto	Neutral
39	AP_ACTCAP	Correcto	Correcto	Incorrecto	Correcto
40	AMC_METODO	Incorrecto	Correcto	Correcto	Correcto
41	AMC_MICROO	Correcto	Correcto	Correcto	Correcto
42	AMC_MEDIO	Correcto	Correcto	Correcto	Correcto
43	ACM_METODO	Correcto	Correcto	Correcto	Correcto
44	ACM_TEMPERATURA	Incorrecto	Correcto	Correcto	Correcto
45	ACM_TIEMPO	Correcto	Correcto	Correcto	Incorrecto

Nota. Esta tabla muestra las etiquetas con los respectivos resultados obtenidos por parte de los modelos.

Con esto se pudo visualizar como actúa el modelo con nuevos artículos científicos, además, con los artículos seleccionados se pudo observar el número de entidades que fueron identificadas como correcto, neutral o incorrecto como se puede observar en la tabla 11 donde

se obtuvieron resultados positivos con una media de 25, 6.5 y 13.5 para los términos correcto, neutral e incorrecto respectivamente.

Tabla 11

Resumen de la comparación de resultados

Términos	X88	X167	X120	X9
Correcto	30	28	10	32
Neutral	3	8	8	7
Incorrecto	12	9	27	6

Nota. Esta tabla muestra un resumen de lo observado en la tabla 10, es decir, el número de entidades que fueron identificadas según el término y su respectivo articulo científico.

Capítulo V

Conclusiones, Recomendaciones y Trabajos Futuros

Conclusiones

En el presente trabajo se desarrollaron dos modelos machine learning y fruto del análisis del estado del arte se determinaron las técnicas necesarias para construir la solución donde todos concuerdan que NER es la mejor opción para encontrar patrones en información. En particular se utilizaron Redes Neuronales Convolucionales y un modelo de transformador pre entrenado denominado RoBERTa que es una versión mejorada de BERT. El uso de ambos modelos proporcionó más datos lo que enriqueció de más información para los investigadores.

El diseño de la arquitectura de la solución se basó en los trabajos relacionados donde mencionaban las fases de preprocesamiento y procesamiento de los datos; para este estudio se descompuso dichas fases en cuatro; la fase de origen de los datos, proceso ETL, entrenamiento y resultados de los modelos.

Respecto al rendimiento de los modelos, estos fueron evaluados con métricas que se utilizan para medir el rendimiento de clasificadores multiclases, es decir, la precisión, recall y puntaje F1. En los resultados de estas evaluaciones se evidencio la necesidad de contar con muchos textos etiquetados pues al tener 45 entidades, lo recomendable seria tener una distribución igualada entre los datos etiquetados de cada entidad y así se obtendrían mucho mejores resultados pues actualmente la precisión promedio de los modelos es de 51,39%.

Por último, la solución se validó por expertos del área de Biotecnología donde se realizó un experimento con artículos científicos que no pertenecían al conjunto de datos utilizados para el entrenamiento y evaluación de los modelos; con esto se contrastaron los resultados obtenidos por los modelos y las anotaciones extraídas manualmente donde se obtuvo una media de 25 entidades extraídas correctamente.

Recomendaciones

Considerando la importancia que tiene esta investigación y en base a los resultados obtenidos, se formulan algunas recomendaciones para continuar con estudios sobre el tema de extracción de textos automatizada, específicamente para el caso de estudio de nanopartículas, pues como se observó en el estado del arte, no existen estudios sobre dicho tema debido a que muchos se enfocan en los campos como la química, medicina, entre otras; además de necesitar un equipo especializado en el tema para obtener anotaciones confiables y de calidad lo que representa tiempo y costos elevados.

Para este proyecto se utilizó un enfoque de aprendizaje supervisado, es decir, se partió de un conjunto de datos etiquetados manualmente lo cual consumió la mayor parte de tiempo de este proyecto. Por esta razón, se recomienda el uso de un enfoque semi supervisado para obtener conjuntos de datos más robustos en menor tiempo y así equilibrar la cantidad de anotaciones por cada entidad; se recomienda realizar un análisis de todas las entidades y eliminar aquellas que sean ambiguas o irrelevantes.

Esta investigación se centró en el desarrollo de los modelos machine learning lo cual impidió la elaboración de una interfaz gráfica que facilite su uso para los investigadores de Biotecnología. Se recomienda desarrollar una aplicación que pueda integrar todo, así como la implementación de una base de datos que almacene los resultados para que los investigadores puedan tener una herramienta de consulta general y no solo una extracción individual de cada artículo científico.

Trabajos Futuros

Este trabajo representa los primeros pasos para la creación de una herramienta automatizada para extraer información de artículos científicos sobre nanopartículas. Sin embargo, aún queda un largo camino por recorrer para conseguir una herramienta que genera

mucha confianza en sus extracciones, por lo que se proponen trabajos que se pueden realizar partiendo de este proyecto.

Como trabajos a futuro se podría realizar un análisis de los modelos que se han desarrollado e investigar cómo podrían ser mejorados, es decir, se podrían probar distintos modelos pues existen varios, por ejemplo, para esta investigación se utilizaron los modelos NER con CNN y RoBERTa.

Además, se podría realizar un análisis con otro caso de estudio donde se realice una comparación entre el modelo construido desde cero y el uso del modelo pre entrenado, en este caso el propuesto en este estudio; y determinar si el modelo pre entrenado es capaz de funcionar en otros modelos que se relacionen al campo de la biotecnología o afines.

Bibliografía

Amador, M. (2009). *¿QUE ES UNA REVISIÓN DE LITERATURA EN UN PROYECTO DE INVESTIGACIÓN?* Recuperado el 18 de Junio de 2022, de METODOLOGIA DE LA INVESTIGACIÓN: https://manuelgalan.blogspot.com/2009/10/que-es-una-revision-de-literatura-en-un.html

Barrios, J. (2022). *Redes Neuronales Convolucionales*. Recuperado el 18 de Junio de 2022, de Big Data: https://www.juanbarrios.com/redes-neurales-convolucionales/

Brocke, J., Maedche, A., & Hevner, A. (2020). Introduction to Design Science Research. *Design Science Research. Cases*, 1-13. doi:10.1007/978-3-030-46781-4_1

Budgen, D., Kitchenham, B., Charters, S., Turner, M., Brereton, P., & Linkman, S. (2007). Preliminary results of a study of the completeness and clarity of structured abstracts. *International Conference on Evaluation and Assessment in Software Engineering (EASE), 11*, 1-9. doi:10.14236/ewic/EASE2007.7

Builtin. (2022). *¿Cómo funciona la inteligencia artificial?* Recuperado el 18 de Junio de 2022, de Builtin: https://builtin.com/artificial-intelligence

Cardellino, F. (2021). *Tutorial de Google BERT para PNL con aprendizaje automático*. Recuperado el 18 de Junio de 2022, de freeCodeCamp: https://www.freecodecamp.org/espanol/news/tutorial-de-google-bert-para-pnl-con-aprendizaje-automatico

Codina, L. (2020). *Estructura y funciones de las bases de datos académicas*. Recuperado el 15 de Marzo de 2022, de lluiscodina.com: https://www.lluiscodina.com/bases-de-datos-academicas-registros/

Codina, L., Morales, A., Rodríguez, R., & Pérez, M. (2020). Uso de Scopus y Web of Science para investigar y evaluar en comunicación social: análisis comparativo y caracterización.

index.comunicación, 10, 235-261. Recuperado el 15 de Marzo de 2022, de https://doi.org/10.33732/ixc/10/03Usodes

Díaz, L., Torruco, U., Martínez, M., & Varela, M. (2013). La entrevista, recurso flexible y dinámico. *Investigación en educación médica*, 162-167.

Explotion AI. (2022). *Training Pipelines & Models*. Recuperado el 21 de Mayo de 2022, de spaCy: https://spacy.io/usage/training#basics

Fonseca, E. (17 de Abril de 2020). Preliminary Literature Review Theory - Video 1 [video]. Youtube. Recuperado el 23 de Marzo de 2022, de https://www.youtube.com/watch?v=3zcY87cV0YQ

Fredes, P. (2017). Herramienta computacional de apoyo a la etapa de búsqueda de estudios primarios en las bases de datos científicas. *(Tesis de ingenieria).* Universidad Catolica de la Santisima Concepcion, Concepción.

Gao, S., Kotevska, O., Sorokine, A., & Christian, J. (2021). A pre-training and self-training approach for biomedical named entity recognition. *PLoS One*, 1-23. doi:https://doi.org/10.1371/journal.pone.0246310

González, A. (2022). *¿Qué es Machine Learning?* Recuperado el 18 de Junio de 2022, de Cleverdata: https://cleverdata.io/que-es-machine-learning-big-data/

Guirao, S. (2015). Utilidad y tipos de revisión de literatura. *SciELO, 9*(2). Recuperado el 18 de Junio de 2022, de https://dx.doi.org/10.4321/S1988-348X2015000200002

Gupta, M. (2018). *A Review of Named Entity Recognition (NER) Using Automatic Summarization of Resumes*. Recuperado el 18 de Junio de 2022, de Towards Data Science: https://towardsdatascience.com/a-review-of-named-entity-recognition-ner-using-automatic-summarization-of-resumes-5248a75de175

Hernando, A. (2019). *¿Qué es y qué no es un artículo científico?* Recuperado el 18 de Junio de 2022, de Revista Comunicar: https://www.revistacomunicar.com/wp/escuela-de-autores/que-es-y-que-no-es-un-articulo-cientifico/

Hong, Z., Tchoua, R., Chard, K., & Foster, I. (2020). SciNER: Extracting Named Entities from Scientific Literature. *Computational Science, 20*, 308-321. doi:https://doi.org/10.1007/978-3-030-50417-5_23

Honnibal, M. (2017). *spaCy's NER model*. Recuperado el 25 de Mayo de 2022, de spaCy: https://spacy.io/universe/project/video-spacys-ner-model#___gatsb

Horev, R. (2018). *BERT Explained: State of the art language model for NLP*. Recuperado el 18 de Junio de 2022, de Towards Data Science: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Ibáñez, A. (2019). *Semi-Supervised Learning…el gran desconocido*. Recuperado el 18 de Junio de 2022, de Telefonica Tech: https://empresas.blogthinkbig.com/semi-supervised-learningel-gran-desconocido/

IBM. (2021). *Acerca de la minería de textos*. Recuperado el 18 de Junio de 2022, de IBM: https://www.ibm.com/docs/es/spss-modeler/18.1.1?topic=analytics-about-text-mining

IBM. (2022). *Machine learning y ciencia de datos*. Recuperado el 18 de Junio de 2022, de IBM: https://www.ibm.com/es-es/analytics/machine-learning

IBM Cloud Education. (2020). *Artificial Intelligence (AI)*. Recuperado el 18 de Junio de 2022, de IBM: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

ICC. (2022). *Big Data & Artificial Inteligence*. Recuperado el 18 de Junio de 2022, de ICC: https://www.iic.uam.es/en/big-data-artificial-inteligence/

Jódar, J. (2010). La era digital: Nuevos Medios, Nuevos Usuarios y Nuevos Profesionales. *Razón y Palabra, 71*, 1-10. Recuperado el 14 de Marzo de 2022, de https://www.redalyc.org/articulo.oa?id=199514914045

Johnson, S. (2014). *What is a Science Article?* Recuperado el 18 de Junio de 2022, de Newton Gresham Library: https://shsulibraryguides.org/c.php?g=86714&p=3217433

Kuriakose, J. (2019). *BIO / IOB Tagged Text to Original Text*. Recuperado el 14 de Mayo de 2022, de Medium: https://medium.com/analytics-vidhya/bio-tagged-text-to-original-text-99b05da6664

Li, Z., Lian, Y., Ma, X., Zhang, X., & Li, C. (2020). Bio-semantic relation extraction with attention-based external knowledge reinforcement. *BMC Bioinformatics, 21*, 1-18. doi:https://doi.org/10.1186/s12859-020-3540-8

Lin, C., Shao, Y., Zhang, J., & Yun, U. (2020). Enhanced Sequence Labeling Based on Latent Variable Conditional Random Fields. *Neurocomputing, 403*, 431-440. doi:https://doi.org/10.1016/j.neucom.2020.04.102

Lorite, A. (2019). *Papers y más papers: las sombras en la industria de las publicaciones científicas*. Recuperado el 15 de Marzo de 2022, de elsaltodiario.com: https://www.elsaltodiario.com/universidad/papers-y-mas-papers-las-sombras-en-la-industria-de-las-publicaciones-cientificas

Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., . . . Rocha, M. (2009). @Note: a workbench for biomedical text mining. *Journal of Biomedical Informatics, 42*, 710-720. doi:https://doi.org/10.1016/j.jbi.2009.04.002

Lutkevich, B. (2021). *Natural language processing (NLP)*. Recuperado el 18 de Junio de 2022, de Tech Target: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

Mahmud, Ö. (2021). *What's ETL?* Recuperado el 13 de Mayo de 2022, de Towards Data Science: https://towardsdatascience.com/whats-etl-b4903a57f8ce

Medina, M., Galván, L., & Reyes, R. (2015). Las nanopartículas y el medio ambiente. *SciELO, 19*. Recuperado el 15 de Marzo de 2022, de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-48212015000100005

Meta AI. (2019). *RoBERTa: An optimized method for pretraining self-supervised NLP systems*. Recuperado el 18 de Junio de 2022, de Meta AI: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/

Morales, A., & Tobar, L. (2020). Sistema Inteligente Basado En Ontologías, Procesamiento Del Lenguaje Natural y Técnicas de Minería de datos para recomendar contenidos Científico-Metodológicos del Ámbito de Terapia del Lenguaje. *(Tesis de ingenieria).* Universidad Politécnica Salesiana, Cuenca.

Moreno, A. (2018). *Procesamiento del lenguaje natural ¿qué es?* Recuperado el 18 de Junio de 2022, de ICC: https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Pardeshi, N. (2020). *Formatting SpaCY custom training data the easier way*. Recuperado el 15 de Mayo de 2022, de Medium: https://medium.com/@nikita25.pardesi/formatting-spacy-custom-training-data-the-easier-way-3aa4f35f6112

Prodigy. (2022). *Named Entity Recognition*. Recuperado el 06 de Junio de 2022, de Prodigy: https://prodi.gy/docs/named-entity-recognition

PubMed. (2022). *Descripción general de PubMed*. Recuperado el 07 de Abril de 2022, de PubMed: https://pubmed.ncbi.nlm.nih.gov/about/

Pykes, C. (2020). *Confusion Matrix "Un-confused"*. Recuperado el 03 de Junio de 2022, de Towards Data Science: https://towardsdatascience.com/confusion-matrix-un-confused-1ba98dee0d7f

Pykes, C. (2021). *The Most Common Evaluation Metrics In NLP*. Recuperado el 02 de Junio de 2022, de Towards Data Science: https://towardsdatascience.com/the-most-common-evaluation-metrics-in-nlp-ced6a763ac8b

Recuero, P. (2022). *Datos de entrenamiento vs datos de test*. Recuperado el 15 de Mayo de 2022, de Telefónica Tech: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/

Russell, J. (2001). La comunicación científica a comienzos del siglo XXI. *Revista Internacional de Ciencias Sociales, 168*, 2-14. Recuperado el 14 de Marzo de 2022

Santos, A. (2022). *¿Como son fabricadas las Nanopartículas?* Recuperado el 18 de Junio de 2022, de Nanovs: https://nanova.org/fabricacion-de-nanoparticulas/

SAS. (2022). *Procesamiento del lenguaje natural*. Recuperado el 18 de Junio de 2022, de SAS: https://www.sas.com/es_ar/insights/analytics/what-is-natural-language-processing-nlp.html#nlpmethods

Sets de Entrenamiento, Test y Validación. (2020). Recuperado el 15 de Mayo de 2022, de Aprende machine learning: https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/

Singh, N. (2020). *Métricas De Evaluación De Modelos En El Aprendizaje Automático*. Recuperado el 02 de Junio de 2022, de DataSource.ai: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Solmeglas Lab. (2020). *¿Qué son las Nanopartículas? Tipos y propiedades de estas partículas*. Recuperado el 18 de Junio de 2022, de Solmeglas Lab: https://solmeglas.com/que-son-nanoparticulas-tipos/

spaCy. (2022). *Training Pipelines & Models*. Recuperado el 26 de Mayo de 2022, de spaCy: https://spacy.io

Swain, M., & Cole, J. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling, 56*, 1894-1904. doi:https://doi.org/10.1021/acs.jcim.6b00207

Universidad de las Fuerzas Armadas ESPE. (2020). *Investigación*. Recuperado el 16 de Marzo de 2022, de investigacion.espe.edu.ec: https://investigacion.espe.edu.ec/estadisticas/

Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., & Rosé, C. (2021). Improving broad-coverage medical entity linking with semantic type prediction and large-scale. *Journal of Biomedical Informatics, 121*, 1-16. doi:https://doi.org/10.1016/j.jbi.2021.103880

Voita, L. (2022). *Convolutional Neural Networks for Text*. Recuperado el 18 de Junio de 2022, de Lena Voita: https://lena-voita.github.io/nlp_course/models/convolutional.html

Wei, Q., Chen, T., Xu, R., He, Y., & Gui, L. (2016). Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *The Journal of Biological Databases and Curation*, 1-8. doi:https://doi.org/10.1093/database/baw140

Apéndices