UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Reconocimiento de entidades nombradas en el dominio farmacéutico

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Marcos Ramón, Mario

Tutor/a: Rosso, Paolo

Cotutor/a externo: ASENSIO MARCO, CESAR

CURSO ACADÉMICO: 2022/2023

Resumen

Actualmente, el Procesamiento del Lenguaje Natural (PLN) y, en concreto, las técnicas de Reconocimiento de Entidades Nombradas (NER) se encuentran en auge, pero la mayor parte de trabajos que se han realizado están enfocados a identificar entidades comunes, como pueden ser personas, organizaciones o localizaciones.

Es por ello que el objetivo del presente trabajo es reconocer entidades que hacen referencia a productos nuevos lanzados por una empresa de una serie de textos en castellano del ámbito farmacéutico, así como clasificar dichos textos en tipo de producto farmacéutico creando una taxonomía previa.

Para conseguir este objetivo, noticias relacionadas con esta temática han sido recuperadas de internet, limpiadas y etiquetadas de manera fiable mediante técnicas de PLN para generar un corpus. A continuación, a varios modelos de lenguaje pre-entrenados se les ha realizado un *fine-tuning* (seleccionar un modelo de lenguaje pre-entrenado y refinar su entrenamiento con un conjunto de datos específico de una tarea en particular), para aprovechar su conocimiento del contexto y el idioma y para resolver las dos tareas mencionadas, es decir, reconocimiento de entidades por un lado y clasificación de textos por otro.

Tras comparar un total de cuatro modelos pre-entrenados (mBERT, BETO, DistilBERT y RoBERTa) y optimizar sus parámetros, el Transformer que mejores resultados ha dado ha sido el basado en BETO, obteniendo para el reconocimiento de las entidades empresa y producto un *Accuracy* de 0.97 y un *F1-score* de 0.67, mientras que para la tarea de clasificación de textos se ha conseguido un 0.83 de *Accuracy* y un 0.82 de *F1-score.*

Palabras clave: Procesamiento del Lenguaje Natural (PLN), Reconocimiento de Entidades Nombradas (NER), ámbito farmacéutico, Transformer, fine-tuning.

Resum

Actualment, el Processament del Llenguatge Natural (PLN) i, en concret, les tècniques de Reconeixement d'Entitats Nomenades (NER) es troben en apogeu, però la major part de treballs que s'han realitzat estan enfocats a identificar entitats comunes, com poden ser persones, organitzacions o localitzacions.

És per això que l'objectiu del present treball és reconèixer entitats que fan referència a productes nous llançats per una empresa d'una sèrie de textos en castellà de l'àmbit farmacèutic, així com classificar aquests textos en tipus de producte farmacèutic creant una taxonomia prèvia.

Per aconseguir aquest objectiu, notícies relacionades amb aquesta temàtica han estat recuperades d'Internet, netejades i etiquetades de manera fiable amb tècniques de PLN per generar un corpus. A continuació, a diversos models de llenguatge pre-entrenats se'ls ha realitzat un *fine-tuning* (seleccionar un model de llenguatge pre-entrenat i refinar el seu entrenament amb un conjunt de dades específic d'una tasca en particular), per aprofitar el coneixement del context i l'idioma i per a resoldre les dues tasques esmentades, és a dir, reconeixement d'entitats per un costat i classificació de textos per un altre.

Després de comparar un total de quatre models pre-entrenats (mBERT, BETO, DistilBERT y RoBERTa) i optimitzar els seus paràmetres, el Transformer que millors resultats ha donat ha estat el basat en BETO, obtenint per al reconeixement de les entitats empresa i producte un *Accuracy* de 0.97 i un *F1-score* de 0.67, mentre que per a la tasca de classificació de textos s'ha aconseguit un 0.83 d'*Accuracy* i un 0.82 de *F1-score*.

Paraules clau: Processament del Llenguatge Natural (PLN), Reconeixement d'Entitats Nomenades (NER), àmbit farmacèutic, Transformer, fine-tuning.

Abstract

Currently, Natural Language Processing (NLP) and, specifically, Named Entity Recognition (NER) techniques are booming, but most of the work that has been done is focused on identifying common entities, such as people, organizations or locations.

That is why the objective of this paper is to recognize entities that refer to new products launched by a company from a series of texts in Spanish in the pharmaceutical field, as well as to classify said texts by type of pharmaceutical product, creating a prior taxonomy.

To achieve this objective, news related to this topic have been retrieved from the internet, cleaned and labeled in a reliable way using NLP techniques to generate a corpus. Next, some pre-trained language models have been fine-tuned (select a pre-trained language model and refine its training with an specific data set of a particular task), to take advantage of their knowledge of the context and the language and to solve the two mentioned tasks, that is, entity recognition on the one hand and text classification on the other.

After comparing four pre-trained models (mBERT, BETO, DistilBERT y RoBERTa) and optimizing their parameters, the Transformer that has given the best results has been the one based on BETO, obtaining an Accuracy of 0.97 and an F1-score of 0.67 for recognition by the company and product entities, while for the text classification task, an Accuracy of 0.83 and an F1-score of 0.82 have been achieved.

Keywords: Natural Language Processing (NLP), Named Entity Recognition (NER), pharmaceutical field, Transformer, fine-tuning.

Reconocimiento de entidades nombradas en el dominio farmacéutico

Índice de contenidos

Reconocimiento de entidades nombradas en el dominio farmacéutico

Índice de figuras y tablas

1. Introducción

La tecnología de Procesamiento del Lenguaje Natural (PLN) [1] es un campo dentro de la Inteligencia Artificial (IA) que busca que las máquinas sean capaces de generar, entender e interpretar el lenguaje humano, para así identificar los elementos más relevantes y las posibles interacciones entre palabras en un texto.

Esta tecnología ha experimentado un gran avance en los últimos años gracias a los grandes volúmenes de datos disponibles, la capacidad de cómputo actual y el desarrollo de modelos de lenguaje basados en Redes Neuronales, especialmente la arquitectura de Transformers. Esta arquitectura aprende a modelar eficazmente las relaciones a largo plazo entre las palabras en el texto mediante el mecanismo de atención, que detecta cómo se influencian y dependen entre sí elementos de datos en una serie.

Gracias a esta casuística, los Transformers han demostrado ser altamente efectivos en una amplia gama de tareas de PLN y problemas de aprendizaje supervisado, desde la traducción automática [2], hasta la generación de textos [3] y la clasificación de estos mismos [4], consiguiendo generar respuestas más precisas y coherentes, como en el caso del Reconocimiento de Entidades Nombradas (NER, por sus siglas en inglés).

Para problemas de Procesamiento del Lenguaje Natural, existen gran cantidad de modelos pre-entrenados, tales como BERT [5] o GPT-3 [6], que tienen una importante capacidad para interpretar el lenguaje e identificar patrones en función del contexto circundante. Pero la mayoría de modelos enfocados en tareas NER se encargan de reconocer palabras que representen lugares, organizaciones, personas, etc.

Es por ello que el presente trabajo intenta ir más allá en el Reconocimiento de Entidades Nombradas y pretende generar un modelo que identifique palabras que pertenezcan a textos relacionados con el lanzamiento de un nuevo producto del ámbito farmacéutico en castellano, además de entrenar otro modelo que sea capaz de clasificar dichos textos por tipo de producto.

1.1. Motivación

El 10 de marzo de 2020, la Organización Mundial de la Salud (OMS) declaró la enfermedad provocada por el virus SARS-CoV-2, mundialmente conocido como COVID-19, una pandemia debido a los altos niveles de propagación y su gravedad [7].

Desde ese momento, empresas farmacéuticas de todo el mundo entraron en la carrera por encontrar una vacuna que inmunizara a la población contra esta enfermedad, por lo que la competencia entre dichas empresas por ver cuál era la primera en alcanzar este objetivo era máxima.

Dentro de este contexto, estar a la última en tecnologías y entender la situación de mercado y el avance del resto de empresas era sustancial para poder alcanzar una vacuna efectiva en el menor tiempo posible.

Es por ello que, aunque actualmente esta necesidad de encontrar una vacuna ya no está presente, muchas otras empresas del sector farmacéutico están interesadas en conocer los productos vendidos por la competencia para mejorar la toma de decisiones interna y desarrollar estrategias más efectivas de comercialización y posicionamiento en el mercado.

Mediante este trabajo y al hilo de esta idea, se pretende generar modelos que identifiquen la empresa y el producto de una serie de noticias relacionadas con lanzamientos de nuevos productos farmacéuticos y que dichos textos se clasifiquen en tipo de producto, para así poder obtener de un vistazo la información relevante y no tener que leer todo el texto.

1.2. Objetivos

Tal y como se ha dicho, los objetivos principales del trabajo son el uso de dos Transformers, uno que sea capaz de identificar entidades, en este caso empresa y producto, de una serie de textos relativos al lanzamiento de un nuevo producto farmacéutico, y otro que consiga clasificar dichas noticias por tipo de producto.

Para poder conseguir estos objetivos principales, hay que cumplir primero una serie de objetivos específicos:

- La recopilación de una serie de artículos que encuadren con el proyecto
- La limpieza de dichos artículos
- La extracción de las entidades que aparecen en el texto para generar un corpus
- La creación de una taxonomía previa que contenga el producto y de qué tipo es
- La adaptación (*fine-tuning*) de un modelo del lenguaje pre-entrenado al dataset generado

1.3. Estructura del TFG

Por otra parte, la estructura de este trabajo está dividida en:

1. **Introducción:** A lo largo de este capítulo se hace una breve introducción del contexto del problema y la motivación para abordarlo, así como de los objetivos y la estructura del presente proyecto.
2. **Estado del arte:** Apartado donde se comentan las técnicas utilizadas para hacer *web scraping*, la complejidad del idioma castellano y las diferentes formas de solucionar un problema de Reconocimiento de Entidades Nombradas (NER) y de Clasificación de textos del dominio farmacéutico.
3. **Análisis del problema:** Este capítulo pretende hacer un análisis tanto de las oportunidades de innovación y negocio como del marco legal y ético para dar lugar a una solución adaptada a nuestro problema a resolver.

4. **Preparación y comprensión de datos:** En este apartado se presentan los datos recopilados y su origen, además del procedimiento utilizado para extraer las entidades nombradas mediante Procesamiento del Lenguaje Natural (PLN) y categorizar cada texto según el tipo de producto que contiene. Asimismo, se aborda la depuración y la adaptación realizada en nuestros datos para facilitar el posterior ajuste al modelo.
5. **Conocimiento extraído y evaluación de los modelos:** Mediante este capítulo se pretenden explicar los pasos realizados para adaptar un modelo de lenguaje pre-entrenado a una tarea en particular, debatir las métricas de evaluación empleadas y seleccionar el Transformer que obtenga mejores resultados para cada uno de los objetivos principales de este trabajo.
6. **Validación y despliegue:** Durante este apartado se validan los resultados del modelo y se comenta el despliegue que se realizará una vez finalice el proyecto.
7. **Conclusiones:** Capítulo donde se resume brevemente el trabajo realizado y las conclusiones obtenidas al respecto.
8. **Trabajos futuros:** Este capítulo contiene una serie de ideas sobre posibles trabajos a realizar en un futuro para mejorar los resultados conseguidos.
9. **Referencias:** Contiene referencias a ciertos enlaces o documentos de interés.

Reconocimiento de entidades nombradas en el dominio farmacéutico

2. Estado del arte

Este capítulo pretende comparar las diferentes técnicas de *web scraping* utilizadas para obtener las noticias del dominio farmacéutico, así como estudiar las características de la lengua castellana y cómo esta se relaciona con PLN.

A continuación, tras establecer una idea inicial de la situación del proyecto, se abordarán algunas de las técnicas aplicadas en el contexto de PLN para extracción de entidades por un lado y para clasificación de textos por otro.

Por último, se explicará en qué consisten los Transformers y se estudiarán los modelos de lenguaje más destacados aplicados al castellano, ya que finalmente se utilizará uno de ellos para dar solución a los problemas expuestos en este trabajo.

2.1. Web scraping

En primer lugar, y con el objetivo de obtener noticias de lanzamientos de nuevos productos farmacéuticos de internet, se ha realizado un estudio para ver qué técnica era la más apropiada para conseguir recuperar dichos textos.

Para ello, y en el contexto actual de infinidad de trabajos realizados con este propósito, se decidió hacer *web scraping*, un proceso de extracción de contenidos y datos de sitios web mediante software, sobre nuestra url de noticias farmacéuticas [30].

En este sentido, tres de las librerías más conocidas y usadas para *web scraping* son BeautifulSoup[1], Scrapy[2] y Selenium[3] [8], de las cuales se han analizado sus ventajas y desventajas para decidir cuál se adapta mejor a nuestro proyecto [9] [10].

- **BeautifulSoup:** Es una librería bastante rápida, fácil de usar e interpretar y resistente a posibles cambios en el frontend, pero solo funciona con documentos HTML y XML ya que extrae la información de páginas web estáticas.
- **Scrapy:** Es la más rápida de todas y puede funcionar con un mayor rango de tipos de documentos, pero es mucho más difícil de usar y de interpretar.
- **Selenium:** Es la más lenta y también es muy complicada de usar e interpretar, además de que puede inducir a errores si se producen cambios en el frontend, pero funciona con muchos tipos de documentos y se adapta a páginas dinámicas.

En definitiva, para el contexto de este problema, parece ser que la librería que mejor se adapta es BeautifulSoup, ya que es la más sencilla a nivel interactivo y otorga una gran rapidez a la hora de hacer *web scraping*.

[1] https://www.crummy.com/software/BeautifulSoup/
[2] https://scrapy.org/
[3] https://www.selenium.dev/

2.2. Complejidad lingüística del castellano

Como ya se ha comentado anteriormente, este trabajo consiste en identificar entidades pertenecientes a textos de dominio farmacéutico y escritos en castellano. Por tanto, es necesario entender la complejidad lingüística del castellano como lengua en comparación con el inglés, lengua de referencia y la más utilizada para tareas de PLN.

Licenciados en filología inglesa defienden que el castellano es mucho más complejo que el inglés [11], tanto a nivel de gramática como de semántica, además de la gran cantidad de formas verbales existentes en castellano y la flexibilidad para construir oraciones, situaciones que no se dan en inglés [12].

Esta conclusión, unida a las diferentes posibilidades en las que puede estar estructurado un texto para anunciar el lanzamiento de un nuevo producto, hacen que la efectividad de las tareas de PLN y el modelo de etiquetado NER resultante puedan verse reducidas. Es por ello que para solucionar este problema pueda ser necesario un etiquetado previo manual o la utilización de algunos modelos pre-entrenados para reconocer ciertas entidades.

2.3. Reconocimiento de Entidades Nombradas (NER) en PLN

El Reconocimiento de Entidades Nombradas (NER) es una tarea bastante compleja dentro del campo del Procesamiento del Lenguaje Natural (PLN), la cual consiste en atender a ciertas normas morfológicas, sintácticas y léxicas para poder identificar en un texto ciertas palabras y etiquetarlas dentro de un conjunto de categorías [13].

Las metodologías empleadas para tareas NER han ido avanzando y mejorando a lo largo del tiempo, desde un enfoque básico basado en reglas, pasando por la introducción de modelos de *Machine Learning*. Ejemplos de ello son la clasificación multiclase, encargada de clasificar cada token (palabra) en un conjunto de categorías predefinidas y sin tener en cuenta el contexto; o los campos aleatorios condicionales (CRF), los cuales extraen de cada token sus *features,* que tratan de representar el contexto para poder identificar las entidades [14].

Actualmente, el Reconocimiento de Entidades Nombradas está focalizado en modelos de *Deep Learning*, especialmente en redes neuronales Bi-LSTM *(Bi-directional Long Short Term Memory)*, que combinan dos redes LSTM, diseñadas para recordar dependencias a largo plazo [15]. Estas redes están formadas por una red que avanza hacia adelante de “derecha a izquierda” y otra que avanza hacia atrás de “izquierda a derecha”, capturando así toda la esencia/contexto de la oración y la información tanto anterior como posterior de cada uno de los tokens [16].

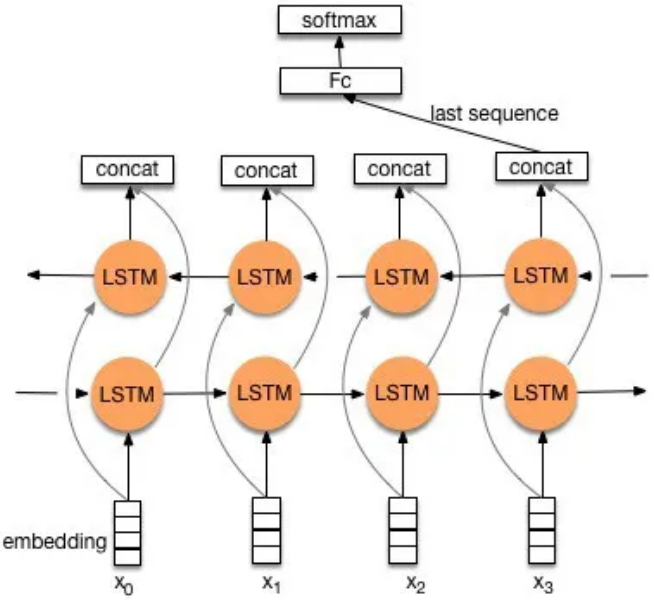


Figura 1. Arquitectura Redes Bi-LSTM [16]

Como se puede ver en la imagen, para cada token representado con la letra x hay una capa de *embedding*, explicada en el apartado 2.5 de este mismo capítulo, capaz de modelar características complejas del uso de las palabras y cómo varían estos usos dependiendo del contexto lingüístico. Así, cada palabra/token estará representado por un único vector de *embedding* y dos palabras similares tendrán vectores de *embedding* similares.

Finalmente, para poder llevar a cabo una tarea NER, es necesario un etiquetado previo de las oraciones de los textos. Uno de los más utilizados es el etiquetado IOB, por el cual cada token tendrá una etiqueta asociada y se tienen en cuenta aquellas entidades formadas por más de una palabra [14]. Para realizar esta codificación, se incluyen los prefijos 'B-' (*beginning*) e 'I-' (*inside*) con el fin de indicar que un token es el inicio de una entidad o pertenece a la misma. Además, la etiqueta 'O' (*outside*) significa que el token no pertenece a ninguna entidad.

Palabra	**Etiquetado IOB**
Pedro	B-PER
Sánchez	I-PER
visitará	O
Reino	B-LOC
Unido	I-LOC
y	O
Bélgica	B-LOC
esta	O
semana	O

Figura 2. Tabla explicativa etiquetado IOB

2.4. Clasificación de textos en PLN

La clasificación de textos es otra de las técnicas más presentes en el Procesamiento del Lenguaje Natural (PLN) y se encarga de asignar un texto a una o más categorías predefinidas con el objetivo de organizar, estructurar y categorizar dicho texto, en este caso perteneciente al dominio farmacéutico [17].

Para abordar esta tarea, se van a estudiar desde clasificadores clásicos como puede ser Naive Bayes, hasta modelos mucho más complejos como las Máquinas de Vectores Soporte (SVM, por sus siglas en inglés) o las Redes Neuronales Recurrentes (RNN, por sus siglas en inglés), entre las que se encuentran las redes LSTM.

2.4.1. Naive Bayes

Este método hace referencia al teorema de Bayes [18], creado por Thomas Bayes y que se basa en hacer la clasificación teniendo en cuenta la probabilidad de ocurrencia de los distintos sucesos. Para ello, se parte de dos sucesos que se consideran independientes (A y B) y se quiere calcular la probabilidad de un determinado suceso en función de la información previa que se tiene del mismo.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Esta es la fórmula de dicho teorema, donde:

- **P(A):** Probabilidad de ocurrencia a priori del suceso A.
- **P(B):** Probabilidad de ocurrencia a priori del suceso B.
- **P(A | B):** Probabilidad de que ocurra A cuando sucede B.
- **P(B | A):** Probabilidad de que ocurra B cuando sucede A.

En definitiva, el modelo de Naive Bayes pretende asignar a un suceso dado su clasificación más probable (probabilidad a posteriori) en base a las clasificaciones de los sucesos anteriores (probabilidad a priori) [19].

2.4.2. Máquinas de Vectores de Soporte (SVM)

Otro clasificador que funciona bastante bien para la clasificación de textos son las Máquinas de Vectores de Soporte (SVM), cuyo objetivo es encontrar el hiperplano óptimo de tamaño n-1 que consigue separar n clases [20].

Para entender mejor el funcionamiento de este clasificador se supone que se tienen 2 clases. Por tanto, el hiperplano será una recta.

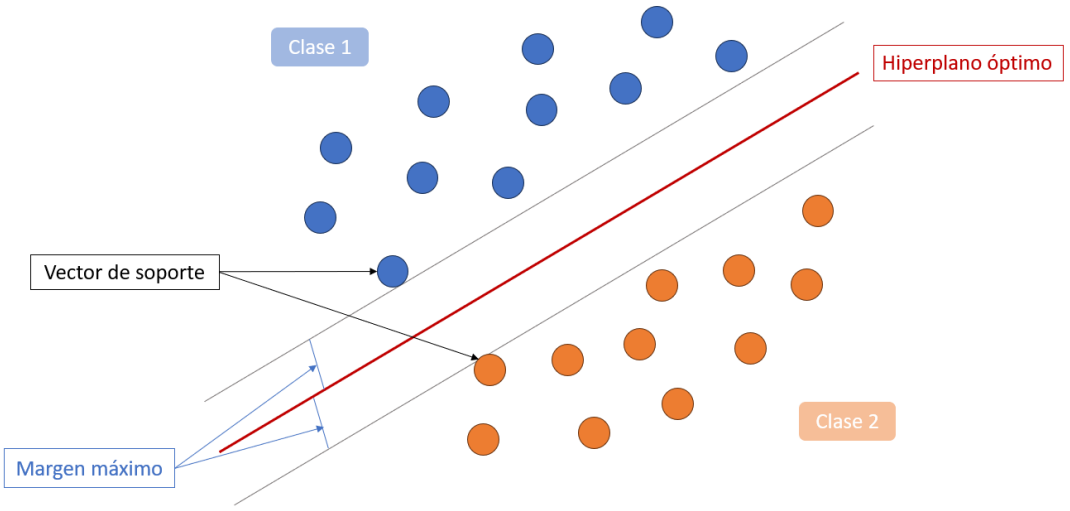


Figura 3. Representación de Máquinas de Vectores Soporte (SVM)

Como se puede ver en la imagen, para obtener el hiperplano óptimo se calcula la distancia perpendicular de cada observación a los hiperplanos, y aquel que obtenga la mayor distancia (margen) para cada clase entre el punto más cercano al hiperplano (vector de soporte) y dicho hiperplano será el óptimo.

Esta definición parece bastante complicada pero, en resumen, el hiperplano óptimo es aquel que se encuentra más alejado de todas las observaciones de entrenamiento, es decir, el que mejor separa las clases.

Por otra parte y como este método fue creado para problemas de 2 clases, existen adaptaciones para trabajar con múltiples clases que tienen como objetivo transformar los datos a binarios, dividiendo un problema multiclase en varios problemas binarios [21]. Estas son:

- **One-vs-one:** Se compara cada una de las clases con cada una de las clases restantes.
- **One-vs-rest:** Se compara cada una de las clases con el resto.

Esta diferencia se ve mejor con un ejemplo. En la Tabla 1 existen 3 clases pertenecientes a tipos de coches (gasolina, diésel e híbrido) donde se comparan ambas adaptaciones.

One-vs-one	One-vs-rest
gasolina vs diésel = diésel vs gasolina	gasolina vs (diésel & híbrido)
gasolina vs híbrido = híbrido vs gasolina	diésel vs (gasolina & híbrido)
híbrido vs diésel = diésel vs híbrido	híbrido vs (gasolina & diesel)

Tabla 1. Comparativa One-vs-one y One-vs-rest

2.4.3. Redes Neuronales Recurrentes (RNN)

Por otro lado, las Redes Neuronales Recurrentes (RNN) son un tipo de redes neuronales que almacenan en memoria la información de las entradas anteriores, para así influir en las entradas y salidas posteriores [22].

Además, una red neuronal recurrente (RNN) es capaz de procesar una secuencia de longitud arbitraria. Para conseguirlo, aplica recursivamente una función de transición f sobre el vector de estado oculto h_t, es decir, para calcular la activación del estado oculto h_t en el paso de tiempo t se aplica la función f sobre la entrada actual x_t y el estado oculto anterior h_{t-1} [23].

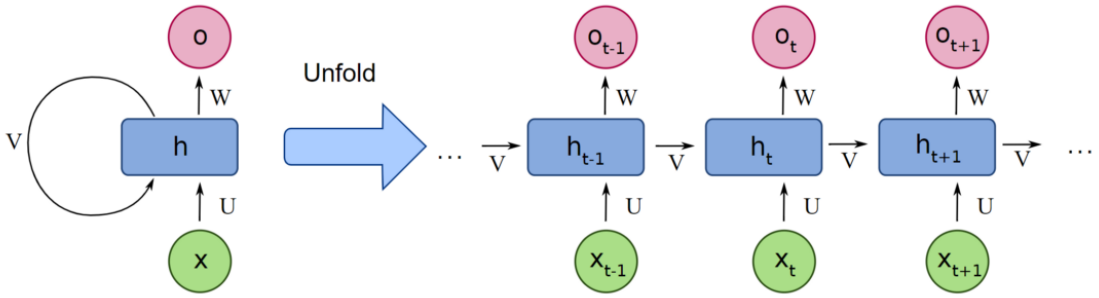


Figura 4. Arquitectura Redes Neuronales Recurrentes (RNN) [34]

En este gráfico se observa la arquitectura de una RNN simple, la cual recibe un input, se aplica una función dada y devuelve un output. Este output junto con el input siguiente serán pasados a la red neuronal y así sucesivamente. Por tanto, cada palabra se procesa individualmente pero teniendo en cuenta el contexto de las palabras anteriores.

Dicho esto, parece ser que esta estructura de RNN es una buena candidata para la tarea de clasificar textos. Sin embargo no es así, ya que durante el entrenamiento, los vectores gradiente (representación matemática que indica la dirección y magnitud del cambio más rápido en una función en un punto dado) pueden crecer o decaer exponencialmente cuando aparecen secuencias extensas de texto, por lo que no funciona bien para frases largas.

Una solución a este problema son las redes LSTM (*Long Short Term Memory*), formadas por un arquitectura que mantiene una celda de memoria separada en su interior que actualiza y expone su contenido sólo cuando se considera necesario [23].

Esta nueva arquitectura proporciona una solución adecuada, pero para textos es importante tener información de todo el contexto, no solo de lo recorrido anteriormente. Es por ello que, al igual que para el apartado anterior de NER, llevar a cabo una arquitectura Bi-LSTM es una muy buena idea, ya que con una red LSTM se recupera toda la información anterior y con la otra red la información posterior.

Con todo lo redactado hasta ahora, una buena arquitectura para Reconocimiento de Entidades Nombradas (NER) es Bi-LSTM, gracias a que almacena información del contexto y durante más tiempo. No obstante, en 2017 se publicó un artículo llamado *"Attention is all you need"* [24], donde se presentaba un nuevo tipo de red neuronal denominada Transformer y que lograba muy buenos resultados para dichas tareas.

2.5. Transformers para PLN

Tal y como hemos comentado en el apartado anterior, en 2017 se estableció una nueva arquitectura de red neuronal llamada Transformer y que revolucionó, entre otros, el mundo del Procesamiento del Lenguaje Natural (PLN). Esta arquitectura fue creada por una serie de expertos en Inteligencia Artificial (IA) y *Deep Learning* y publicada en el artículo *"Attention is all you need"* [24].

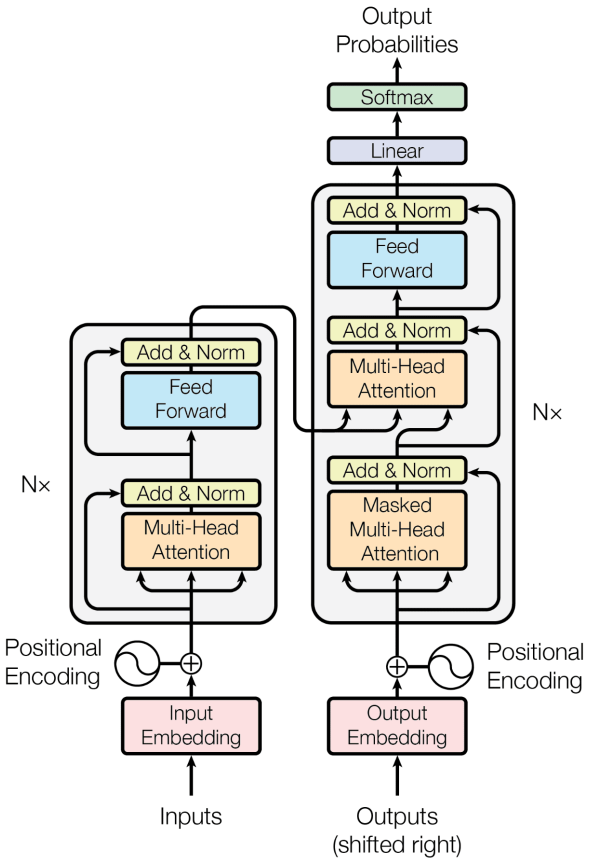


Figura 5. Arquitectura Transformers [24]

Como se puede ver en la imagen, la arquitectura de los Transformers está dividida en dos secciones, el *Encoder* (codificador) a la izquierda y el *Decoder* (decodificador) a la derecha. Pero antes de explicar estas secciones, vamos a definir una serie de conceptos necesarios para entender el funcionamiento de los Transformers:

- ***Word embedding*:** Es una técnica que consiste en representar palabras con vectores de números, denominados vectores de *embedding*, que sirven para crear una representación más semántica de cada palabra, donde palabras semánticamente parecidas se encontrarán cerca en dicho espacio vectorial [50].
- ***Positional encoding*:** Como el modelo no contiene recurrencia ni convolución, más bien, no tiene forma de saber la posición de las palabras, añadimos información extra de posición al vector de *embedding* mediante esta técnica.

- ***Self-attention*:** Es un mecanismo que permite al modelo mirar otras posiciones en el vector de entrada, es decir sobre sí mismo, en busca de posibles relaciones con la palabra actual, y genera un vector basado en las dependencias de cada palabra con el resto de las palabras de la oración. Para ello existen tres vectores (*query, key* y *value*) que generan un output, calculado como una suma ponderada de los valores (*values)*, donde el peso asignado a cada valor (*value)* se calcula mediante una función de compatibilidad de la *query* con la *key* correspondiente.
- ***Multi-Head attention*:** es un tipo de *Self-attention* que contiene varios cabezales y cuyo objetivo es permitir que el modelo atienda a la vez la información de diferentes subespacios de representación en diferentes posiciones.

Tras estas definiciones, vamos a explicar las dos partes principales de la arquitectura Transformer:

Por un lado, el Codificador (*Encoder*) se encarga de procesar la entrada y extraer representaciones de palabras contextualizadas mediante el mecanismo de atención. Cada palabra se representa como un vector contextual que captura su significado en función de su contexto en la oración.

Por otro lado, el Decodificador (*Decoder*) recibe la representación del codificador como entrada y, en función de ella, genera una secuencia de salida, token por token. En esta sección es donde se especializa el Transformer para dar solución a una tarea específica.

En definitiva, los Transformers han implementado mejoras sustanciales con respecto a las redes RNN, entre las que se encuentran la reducción de la complejidad computacional total por capa y el aumento de la cantidad de computación que se puede paralelizar. Estas mejoras, unidas a la posibilidad de hacer *fine-tuning* sobre modelos pre-entrenados para adaptarlos al problema que se desee resolver, hacen que los Transformers sean la arquitectura idónea para alcanzar los objetivos de este proyecto.

Dicho esto, modelos pre-entrenados como la versión multilingüe de BERT (mBERT) [27] o BETO (BERT para el castellano) [25] han obtenido muy buenos resultados para problemas de Procesamiento del Lenguaje Natural en castellano, tal y como se puede ver en la Tabla 2. En esta tabla se comparan varias tareas realizadas sobre diferentes conjuntos de datos, donde destaca NER-C ya que es la más relacionada con los objetivos del trabajo.

Task	BETO-cased	BETO-uncased	Best Multilingual BERT	Other results
POS	**98.97**	98.44	97.10 [2]	98.91 [6], 96.71 [3]
NER-C	**88.43**	82.67	87.38 [2]	87.18 [3]
MLDoc	95.60	**96.12**	95.70 [2]	88.75 [4]
PAWS-X	89.05	89.55	90.70 [8]	
XNLI	**82.01**	80.15	78.50 [2]	80.80 [5], 77.80 [1], 73.15 [4]

Tabla 2. Resultados mBERT vs BETO para PLN en castellano [25]

2.5.1. BERT

BERT *(Bidirectional Encoder Representations from Transformers)* es un modelo de lenguaje desarrollado por *Google AI* en 2018 que ha tenido un gran impacto en el campo del Procesamiento del Lenguaje Natural (PLN) [5]. Es una arquitectura de red neuronal basada en la familia de modelos de atención llamada Transformers.

A diferencia de los modelos de lenguaje tradicionales, que solo ven las palabras en una dirección (izquierda a derecha o derecha a izquierda), BERT procesa el texto en ambos sentidos simultáneamente. Esto permite que el modelo tenga una comprensión más profunda de las palabras y su significado en el contexto en el que aparecen.

Además, este modelo de lenguaje ha sido pre-entrenado en grandes conjuntos de datos sin etiquetar y siguiendo un proceso que consta de dos tareas principales:

- ***Masked Language Model* (MLM):** donde cierta proporción de las palabras de entrada se seleccionan aleatoriamente y se enmascaran (se reemplazan con un token especial "[MASK]"). Luego, el modelo intenta predecir las palabras enmascaradas utilizando el contexto de las palabras circundantes. El objetivo es que el modelo aprenda a capturar las relaciones entre las palabras y las características contextuales para hacer predicciones precisas de las palabras ocultas.
- ***Next Sentence Prediction* (NSP):** donde se tienen pares de oraciones, y el modelo debe predecir si la segunda oración sigue a la primera en el texto original. Esta tarea ayuda al modelo a aprender a capturar información de dependencia entre oraciones y comprender la estructura secuencial del texto.

En general, estas dos tareas en las que se basa el pre-entrenamiento de BERT son esenciales, ya que permiten al modelo aprender representaciones semánticas ricas y contextuales a partir de grandes cantidades de datos sin etiquetar.

Después del pre-entrenamiento, BERT se puede ajustar o *"fine-tunear"* para tareas específicas de Procesamiento del Lenguaje Natural, tales como clasificación de texto, extracción de información, etc., utilizando conjuntos de datos etiquetados más pequeños y específicos para estas tareas [28].

Finalmente, cabe destacar que, como los artículos están escritos en castellano, es necesario el uso de un BERT multilingüe que ha sido entrenado también para el castellano, ya que inicialmente fue entrenado sólo para textos en inglés [27]. Este modelo multilingüe está entrenado con textos en minúsculas y recibe como entrada oraciones separadas por unos tokens especiales de inicio y separación.

Figura 6. Estructura oraciones BERT

2.5.2. BETO

Por otra parte, en 2019 se crea BETO [25], un modelo del lenguaje en castellano basado en Transformers, el cual es, en resumen, la versión en español del BERT original, pero en este caso teniendo en cuenta la acentuación y la letra “ñ”.

Este modelo fue entrenado con gran cantidad de textos en castellano extraídos de diferentes organizaciones como Wikipedia, Naciones Unidas o DGT, y al igual que en BERT, se puede aplicar *fine-tuning* sobre este modelo para adecuarlo a una tarea en específico.

En definitiva, tanto el modelo BERT multilingüe en español como BETO funcionan de una manera muy similar, lo único que cambia de uno a otro son los datos con los que han sido entrenados y que pueden afectar en cierta medida a los resultados obtenidos, ya que el primero es un modelo general adaptado y traducido al castellano, mientras que el segundo ha sido entrenado específicamente para el español.

3. Análisis del problema

Una vez descrito brevemente el Estado del arte, vamos a realizar el análisis del problema a resolver, explicando el contexto en el que se encuentra y el conjunto de soluciones posibles planteadas hasta llegar a la solución propuesta.

En primer lugar, un trabajo relacionado con Transformers en el ámbito farmacéutico presenta diversas oportunidades de innovación y negocio. Algunas de ellas son:

- **Mejora de la eficiencia en el procesamiento de información:** Los Transformers son modelos de aprendizaje automático altamente eficientes para procesar grandes cantidades de datos y extraer información relevante. Aplicar esta tecnología en el ámbito farmacéutico puede agilizar la interpretación de datos de nuevos lanzamientos de fármacos, así como la identificación de patrones de tipos de producto y empresa o la agrupación de empresas por nichos de mercado, lo que puede conducir a una toma de decisiones más rápida y precisa.

- **Análisis de competidores:** Mediante la identificación de entidades farmacéuticas en los textos de lanzamiento de productos, se pueden obtener *insights* valiosos sobre las estrategias y productos de los competidores. Esto podría permitir a las empresas farmacéuticas identificar oportunidades para diferenciarse y desarrollar estrategias más efectivas de comercialización y posicionamiento en el mercado.

- **Mejora en la toma de decisiones:** El uso de este tipo de Transformers puede proporcionar una base sólida de información para la toma de decisiones en el ámbito de la investigación, desarrollo y comercialización de nuevos productos farmacéuticos. Esto puede ayudar a las empresas a tomar decisiones informadas y estratégicas sobre el desarrollo de productos y la asignación de recursos.

Estas son algunas de las oportunidades de innovación y negocio que se pueden explorar en un TFG sobre Transformers en el dominio farmacéutico. Como se ha visto, la combinación de la capacidad de procesamiento de información de los Transformers con el conocimiento y la experiencia en el campo farmacéutico puede ayudar a mejorar los procesos empresariales con el desarrollo de productos muy demandados y a aumentar las ventas de las organizaciones.

No obstante, la tarea que se quiere llevar a cabo con este trabajo es muy específica y actualmente no existe ningún Transformer que sea capaz de identificar entidades como empresa y producto dentro de un texto en castellano referente al lanzamiento de un nuevo producto farmacéutico. Esta casuística complica el trabajo y los resultados obtenidos puede que no sean óptimos, pero si los resultados son satisfactorios, este proyecto puede ayudar a las empresas farmacéuticas a conocer el contexto en el que se encuentran e idear estrategias de negocio adaptadas a dicho contexto.

Por otro lado, y antes de explicar la solución propuesta, se va a estudiar el marco legal y ético en el que se encuentra el proyecto.

3.1. Marco legal y ético

Con todo lo comentado anteriormente, se conoce que los datos se recopilan de internet, en concreto de la página web www.farmaventas.es, y que se basan en noticias de lanzamientos de nuevos productos por parte de empresas pertenecientes al ámbito farmacéutico. Es por ello que dichas noticias no contienen datos sensibles, personales, ni discriminatorios hacia ciertos grupos de personas o con cierto grado de sentimiento, son textos neutrales que no deben ser anonimizados ni tratados de forma especial, y que se almacenan tal y como se recuperan de la web.

Por otro lado, es importante respetar los derechos de autor y la propiedad intelectual de la persona o personas que redactan estas noticias, así como cumplir con la normativa y regulación de la página web. A este respecto, dentro de los términos y condiciones de la web y su política de privacidad [29], se comenta que:

- *'Los usuarios de Internet que accedan a esta WEB pueden visualizar la información contenida en la misma y efectuar downloads o reproducciones privadas en su sistema informático, siempre que los elementos reproducidos no sean cedidos posteriormente a terceros'*

- *'El usuario, deberá utilizar los contenidos e informaciones recogidos en LAS WEBS de forma diligente, correcta y lícita, y en concreto, únicamente para uso personal y no comercial, siempre y cuando no se elimine o modifique el contenido o cualquier mención de fuentes, copyright y demás datos identificativos de derechos de PODIUM GLOBAL MEDIA, S.L. o de terceros, es decir respetando su forma original.'*

En ese sentido, el único propósito del presente trabajo es recuperar las noticias de esta página web y almacenarlas en una base de datos privada que se utilizará para *'fine-tunear'* el Transformer, concepto explicado en el apartado 5.1 del capítulo 5, y dicho Transformer será el que posteriormente recibirá un texto cualquiera relacionado con el lanzamiento de un nuevo producto farmacéutico (no tiene porque ser obtenido de www.farmaventas.es), e identificará la empresa y el producto contenidos en dicho texto y el tipo de producto al que hace referencia.

Por consiguiente, el contenido de dicha página web sólo será utilizado para *'fine-tunear'* al Transformer y el dataset con las noticias recuperadas no será accesible a nadie externo. Además, cabe destacar que este proyecto es meramente educativo y divulgativo, no existe ningún fin lucrativo ni comercial, por tanto, se cumple con la normativa de privacidad de la página web y se puede hacer *web scraping* sobre la misma para recuperar las noticias necesarias.

Finalmente, es importante cumplir con los principios éticos a la hora de recuperar noticias de internet o utilizar artículos de investigación publicados por otros autores. Es por ello que todas aquellas fuentes y contenidos empleados están adecuadamente mencionados y referenciados a lo largo de todo el documento en formato APA.

3.2. Solución propuesta

Con las opciones barajadas en el capítulo de Estado del arte para resolver este problema, la decisión recae en utilizar Transformers, ya que es una arquitectura que está a la orden del día y la que actualmente obtiene mejores resultados para las dos tareas a desarrollar.

Dicho esto, se va a explicar cómo funciona un transformer para Reconocimiento de Entidades Nombradas (NER) por un lado, y para clasificación de textos por otro, con un ejemplo basado en BERT (sirve también para BETO ya que tienen la misma estructura), debido a que, como se ha comentado anteriormente, son modelos lingüísticos basados en redes neuronales usados para tareas de Procesamiento del Lenguaje Natural.

3.2.1. Reconocimiento de Entidades Nombradas (NER)

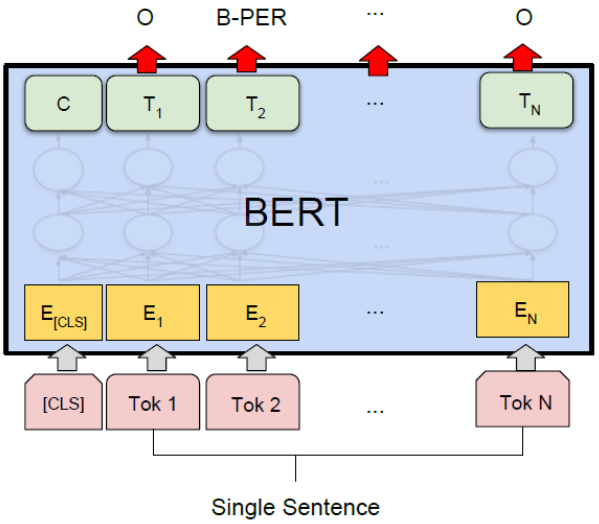


Figura 7. BERT aplicado a Reconocimiento de Entidades Nombradas (NER) [26]

Como se puede ver en la imagen anterior, BERT en el contexto de una tarea NER recibe una única frase dividida en tokens cuyo inicio se identifica con el comando CLS y que es pasada como entrada al Transformer. Los tokens de dicha frase son convertidos en vectores de *embedding*, y tras los procesos realizados por parte del *Encoder* y el *Decoder*, se devuelve la entidad perteneciente a cada token en formato IOB.

3.2.2. Clasificación de textos

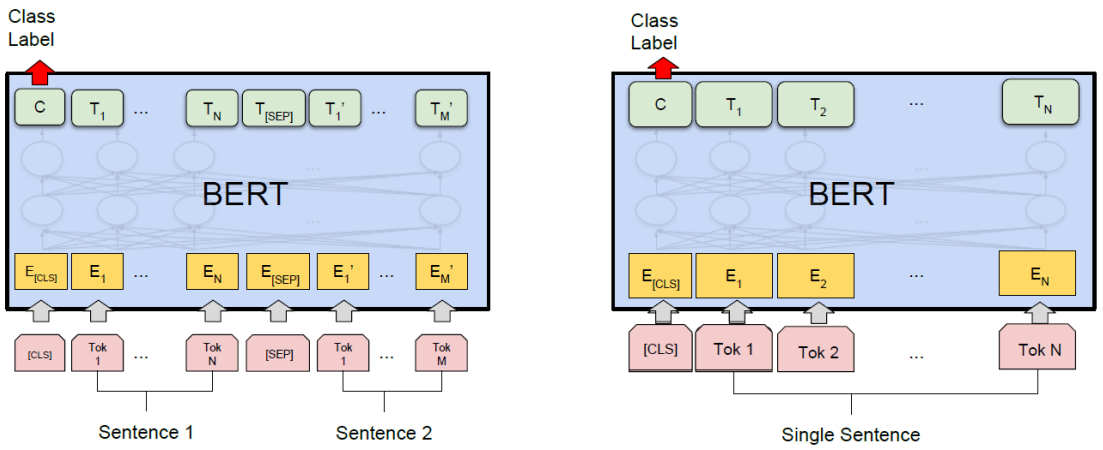


Figura 8. BERT aplicado a clasificación de textos [26]

Por otra parte, para la clasificación de textos, BERT puede procesar una o dos frases a la vez tal y como se observa en la figura superior, donde el inicio vendrá representado por el comando CLS y la separación entre frases por el comando SEP. Dicho esto, el funcionamiento es igual que para NER, donde los tokens de la oración u oraciones son transformados en vectores de *embedding* y procesados por el Transformer, pero en este caso la salida del Transformer es la clase a la que pertenece la frase o frases, dependiendo de la situación.

En resumen, la misión del presente trabajo es ejecutar un fine-tuning de dos Transformers basados en modelos del lenguaje pre-entrenados, uno para Reconocimiento de Entidades Nombradas (NER) y otro para clasificación de textos.

4. Preparación y Comprensión de Datos

En este capítulo se va a realizar una introducción a los datos recogidos y su procedencia, así como el proceso llevado a cabo para extraer las entidades nombradas mediante PLN y clasificar cada uno de los textos en función del tipo de producto que aparece en dicho texto. Además, se comentará también la limpieza y la adaptación realizada sobre los datos para poder ejecutar el *fine-tuning* posterior.

Primeramente, la web de la cual se obtienen los datos es www.farmaventas.es, y en especial aquellos artículos que hacen referencia al lanzamiento de un nuevo producto farmacéutico [30]. Para conseguirlo, tal y como se ha mencionado anteriormente, se utiliza la librería BeautifulSoup de Python[1] para hacer *web scraping* sobre dicha url y obtener el texto de cada una de las noticias.

Para almacenar dichos textos de forma limpia y estructurada, se lleva a cabo una normalización, por la cual se eliminan los acentos y ciertos signos de puntuación, como las exclamaciones o las interrogaciones, y se cambia la letra 'ñ' por la 'n'. Además, debido a que la web se trata como un HTML para que pueda ser procesada por BeautifulSoup, hay ciertos comandos como '<p>' o '' o etiquetas como '\n', '\r' o '\t' que son recuperadas tras hacer *web scraping* y borradas con la librería re[2] de Python referente a operaciones con expresiones regulares.

Una vez hecha la limpieza, la siguiente tarea es la identificación automática de la empresa y el producto en cada uno de los textos de dominio farmacéutico. Un ejemplo del objetivo a conseguir es el siguiente, donde la empresa está subrayada en verde y el producto en amarillo.

Figura 9. Detección de producto y empresa separados por verbo [35]

Por consiguiente, para reconocer la empresa, se ha empleado la biblioteca HuggingFace Transformers[3], que contiene una gran cantidad de modelos pre-entrenados, de los que se ha seleccionado un modelo multilingüe basado en DistilBERT para la tarea NER [31], que ha obtenido buenos resultados en la identificación de ciertas entidades, entre ellas organizaciones, por lo que puede utilizarse para este cometido.

No obstante, de este modo no se conseguía identificar más del 20% de las empresas de forma automática, por lo que se decidió usar otra librería llamada stanza[4], que se encarga de realizar un análisis sintáctico de las oraciones y las palabras. Con este contexto, se identifica el primer verbo del título del artículo, ya que casi siempre son palabras como 'lanza', 'estrena' o 'presenta', y se elige como empresa las palabras que aparecen antes del verbo.

[1] https://www.python.org/
[2] https://docs.python.org/es/3/library/re.html
[3] https://huggingface.co/docs/transformers/index
[4] https://github.com/stanfordnlp/stanza

Por otro lado, para identificar el producto, después de analizar manualmente ejemplos del dataset, se buscará la raíz 'nuev' en el título de cada texto, y de los títulos con este formato, se seleccionará como producto las palabras después del término con raíz 'nuev' y que se encuentren antes de la preposición 'de'.

Además, con este formato se puede identificar también la empresa de forma automática, que, cómo se puede ver en el ejemplo siguiente, son las palabras después de la preposición 'de' (como en el ejemplo anterior, la empresa está subrayada en verde y el producto en amarillo).

Figura 10. Detección de producto por raíz 'nuev' y empresa después de 'de' [36]

Si de esta forma no se consigue reconocer el producto, se empleará de nuevo la librería stanza para identificar el verbo del título. Una vez identificado, así como se puede ver ejemplificado en la Figura 9, el producto estará formado por aquellas palabras que aparezcan después del verbo.

Tras la explicación de la tarea de reconocimiento de entidades, se recuperan un total de 2168 noticias de la página web, de las cuáles hemos borrado 43 porque no trataban sobre el lanzamiento de un nuevo producto.

Agua de mar para prevenir las alergias otoñales.Aunque la primavera es la época más conocida por sus reacciones alérgicas, también en el otoño se produce la proliferación de algunos de los alérgenos más importantes debido a las lluvias y a la humedad en e

Una higiene nasal correcta, practicada con regularidad a base de agua de mar, aporta numerosos beneficios para la salud. Además de reforzar una función defensiva de las mucosas respiratorias y mejorar el drenaje lacrimal, previene de resfriados, de gripe, y reduce la respuesta alérgica por desensibilización de las mucosas.

Paralelamente, es básico aumentar las defensas para prevenir los síntomas de alergia, a través de una **dieta rica y variada**, en la que no falten alimentos ricos en vitaminas y sales minerales. La **hidratación**, siempre es importante, pero en temporada de alergia es recomendable hidratarse durante todo el día, a partir de suplementos a base de agua de mar, que contienen los **minerales** que el organismo necesita.

"El agua de mar hace que las células beban y restauren su morfología. Para restablecer internamente las células y por tanto los órganos, es muy recomendable comenzar durante el verano un tratamiento basado en **agua de mar,** que tiene un efecto drenante e inmunoestimulante, provocando una acción natural del organismo frente a las agresiones diarias en el otoño", según el Dr. **Marco Francisco Payá**, doctor en Medicina por la Universidad de Montpellier (Francia), diplomado en Medicina Manual y Osteopatía, Homeopatía y Fitoterapia y director médico de **Laboratorios Quinton.**

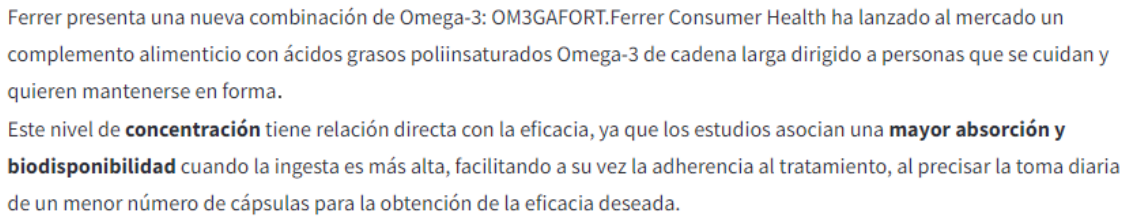
Figura 11. Ejemplo noticia que no trata sobre el lanzamiento de un nuevo producto [37]

En consecuencia, el total de las noticias después del borrado es de 2125, y de éstas se han identificado de forma automática la cantidad de entidades que se ve en la tabla siguiente y el etiquetado de las restantes se ha realizado a mano.

	Empresa	Producto
Entidades identificadas automáticamente	988 entidades	266 entidades
Entidades identificadas manualmente	1137 entidades	1859 entidades

Tabla 3. Entidades identificadas automáticamente y manualmente

Observando la gran cantidad de entidades que se han identificado manualmente, es importante comentar que estos valores son tan altos porque muchas veces la empresa o el producto aparecían escritos de diferente forma a lo largo del texto en comparación a como estaban escritos en el título.

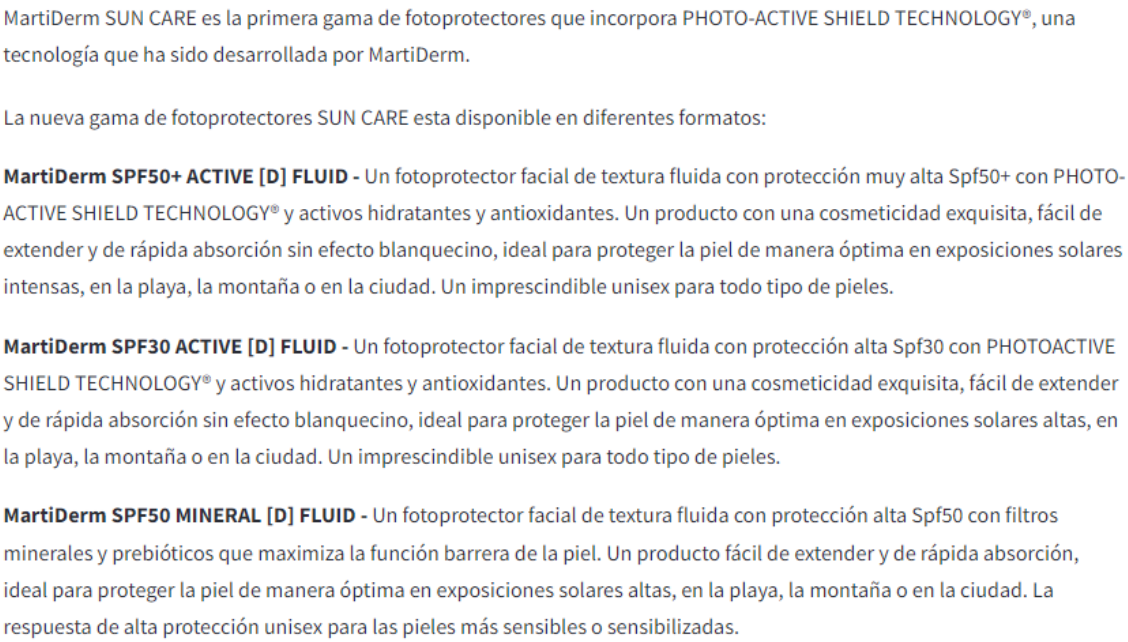


Ferrer presenta una nueva combinación de Omega-3: OM3GAFORT.Ferrer Consumer Health ha lanzado al mercado un complemento alimenticio con ácidos grasos poliinsaturados Omega-3 de cadena larga dirigido a personas que se cuidan y quieren mantenerse en forma.

Este nivel de **concentración** tiene relación directa con la eficacia, ya que los estudios asocian una **mayor absorción y biodisponibilidad** cuando la ingesta es más alta, facilitando a su vez la adherencia al tratamiento, al precisar la toma diaria de un menor número de cápsulas para la obtención de la eficacia deseada.

Figura 12. Ejemplo noticia con empresa escrita de varias formas [38]

O en el título sólo aparecía la gama de productos que se lanzaba y en el cuerpo de la noticia se hablaba sobre los productos que contenía esa gama.



MartiDerm SUN CARE es la primera gama de fotoprotectores que incorpora PHOTO-ACTIVE SHIELD TECHNOLOGY®, una tecnología que ha sido desarrollada por MartiDerm.

La nueva gama de fotoprotectores SUN CARE esta disponible en diferentes formatos:

MartiDerm SPF50+ ACTIVE [D] FLUID - Un fotoprotector facial de textura fluida con protección muy alta Spf50+ con PHOTO-ACTIVE SHIELD TECHNOLOGY® y activos hidratantes y antioxidantes. Un producto con una cosmeticidad exquisita, fácil de extender y de rápida absorción sin efecto blanquecino, ideal para proteger la piel de manera óptima en exposiciones solares intensas, en la playa, la montaña o en la ciudad. Un imprescindible unisex para todo tipo de pieles.

MartiDerm SPF30 ACTIVE [D] FLUID - Un fotoprotector facial de textura fluida con protección alta Spf30 con PHOTOACTIVE SHIELD TECHNOLOGY® y activos hidratantes y antioxidantes. Un producto con una cosmeticidad exquisita, fácil de extender y de rápida absorción sin efecto blanquecino, ideal para proteger la piel de manera óptima en exposiciones solares altas, en la playa, la montaña o en la ciudad. Un imprescindible unisex para todo tipo de pieles.

MartiDerm SPF50 MINERAL [D] FLUID - Un fotoprotector facial de textura fluida con protección alta Spf50 con filtros minerales y prebióticos que maximiza la función barrera de la piel. Un producto fácil de extender y de rápida absorción, ideal para proteger la piel de manera óptima en exposiciones solares altas, en la playa, la montaña o en la ciudad. La respuesta de alta protección unisex para las pieles más sensibles o sensibilizadas.

Figura 13. Ejemplo noticia de lanzamiento de una gama de productos [39]

Además de muchas otras casuísticas, ya que se están tratando textos escritos en castellano, idioma que da lugar a diferentes estructuras gramaticales y semánticas y es muy flexible a la hora de construir oraciones.

Tras este etiquetado, el siguiente objetivo es crear una taxonomía por tipo de producto para clasificar los textos. Esta clasificación está basada en lo publicado en un blog farmacéutico del mayor grupo a nivel nacional de escuelas online [32], dónde se diferencian los productos farmacéuticos en dos grupos principales: medicamentos y cosméticos. A su vez, este último está dividido en subgrupos, como son productos para la piel, para el cabello, etc.

Con toda esta información y en el contexto del problema, la taxonomía previa creada está dividida en 5 clases. Para las 4 primeras, se busca la aparición de ciertas palabras clave en el texto que hagan referencia a una de las clases, mientras que los textos que no contienen ninguna de estas palabras se clasifican como 'Otros'.

Clase	Palabras clave
Productos para la piel	'piel', 'acne', 'maquill'
Productos para el cabello	'pelo', 'cabello'
Medicamentos / complementos	'medica', 'complement'
Perfumes / desodorantes	'perfum', 'colonia', 'desodorant', 'olor'
Otros	Ninguna de las palabras anteriores

Tabla 4. Clases por tipo de producto y sus respectivas palabras clave

Con esta forma de asignar automáticamente las clases a los textos que hagan referencia, se han clasificado de forma correcta 1811 textos, por lo que los 314 restantes se han etiquetado manualmente. En la tabla siguiente se pueden observar la cantidad de noticias existentes para cada una de las clasificaciones, dónde hay 40 noticias en dos clases, 'Productos para la piel' y 'Productos para el cabello', que contienen productos de ambos tipos y por consiguiente serán analizadas dos veces, una por clase.

Clase	Cantidad
Productos para la piel	1198 noticias
Productos para el cabello	177 noticias
Medicamentos / complementos	372 noticias
Perfumes / desodorantes	31 noticias
Otros	387 noticias

Tabla 5. Cantidad de noticias por clase

Una vez se tengan todos los textos con las entidades nombradas reconocidas y las clasificaciones hechas, se transformará dicho texto para que sea procesado posteriormente por un Transformer. Para ello y como se ha comentado anteriormente, se llevará a cabo un etiquetado IOB con un previo etiquetado básico para la tarea de NER.

Este etiquetado básico consiste en, para cada uno de los textos, en identificar dónde se encuentra el token que hace referencia a una entidad nombrada, en nuestro caso a empresa o producto, poner entre corchetes dicho token, y a su lado y entre paréntesis si se trata de una empresa o un producto.

A continuación, dicho etiquetado básico será transformado en un etiquetado IOB, donde para cada texto, se buscan los tokens que están entre corchetes y el tipo de token entre paréntesis, y se le asigna una 'B-' a la primera palabra del token y una 'I-' al resto de palabras del token. Además, a aquellas palabras que no están entre corchetes se les asigna una 'O'.

Por tanto, tal y como se observa en la Tabla 6, el resultado son cinco tokens en formato IOB, los cuales se transforman a números para ser entrenados por el Transformer para NER.

Token IOB	Número
O	0
B-empresa	1
B-producto	2
I-empresa	3
I-producto	4

Tabla 6. Transformación tokens IOB a números

En el ejemplo siguiente se puede ver de forma más clara y concisa cómo funcionan estas permutaciones.

TEXTO NORMAL:

laboratorios babe amplia la linea super fluid, que esta primavera da la bienvenida a un nuevo producto: super fluid depigment

ETIQUETADO BÁSICO:

[laboratorios babe](empresa) amplia la linea super fluid, que esta primavera da la bienvenida a un nuevo producto: [super fluid depigment](producto)

ETIQUETADO IOB:

[B-empresa , I-empresa , O , O , O , O , O , O , O , O , O , O , O , O , O , O , O , B-producto , I-producto , I-producto]

TRANSFORMACIÓN NUMÉRICA:

[1 , 3 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 0 , 2 , 4 , 4]

Figura 14. Ejemplo transformación texto normal a entrada para Transformer [40]

Además, es importante mencionar que a las oraciones se les ha aplicado un tratamiento basado en *subwords*, consistente en dividir las palabras en unidades más pequeñas, lo que permite al modelo manejar palabras desconocidas o raras, ser más generalista y tener un tamaño de vocabulario más reducido. A estas *subwords* generadas, al igual que para los tokens especiales de inicio y separación de la frase ([CLS] y [SEP] respectivamente), se les aplica el número -100. Cuando este número esté presente, se pondrá un 0 en la capa de *attention_mask* para que el Transformer no tenga en cuenta estos tokens y no atienda sus posiciones en el mecanismo de atención.

Por otro lado, para poder adaptar las cinco clases generadas a la arquitectura de Hugging Face para Transformers [33], se deben pasar a números dichas clases y asociar los números a sus textos correspondientes.

Clase	Número
Productos para la piel	0
Productos para el cabello	1
Medicamentos / complementos	2
Perfumes / desodorantes	3
Otros	4

Tabla 7. Transformación clases a números

Tras todos estos procesos, los datos ya están preparados para hacer *fine-tuning* sobre un modelo de lenguaje, con el objetivo de poder entrenar ambos transformers y obtener una solución para nuestro problema.

5. Conocimiento Extraído y Evaluación de Modelos

En este capítulo se van a desarrollar las técnicas y modelos utilizados para generar dos Transformers que sean capaces de solucionar los dos problemas comentados a lo largo del presente trabajo.

Para ello, en primer lugar se explicará en qué consiste el *fine-tuning* y los pasos realizados para adaptar un modelo de lenguaje pre-entrenado a una tarea en particular, a continuación se detallarán las métricas de evaluación empleadas para, finalmente, comparar una serie de modelos '*fine-tuneados'* y seleccionar aquel modelo que mejores resultados proporcione.

5.1. Fine-tuning

Una vez los datos están preprocesados y limpios, con los textos en formato IOB y las clasificaciones de dichos textos convertidas a número entero, se va a realizar un proceso de *fine-tuning* sobre un modelo de lenguaje aplicable al castellano.

Este proceso consiste en tomar un modelo de lenguaje pre-entrenado y luego refinar su entrenamiento con un conjunto de datos específico de una tarea particular. En lugar de entrenar el modelo desde cero, el *fine-tuning* aprovecha el conocimiento y la capacidad de generalización del modelo pre-entrenado y lo adapta a una tarea más específica.

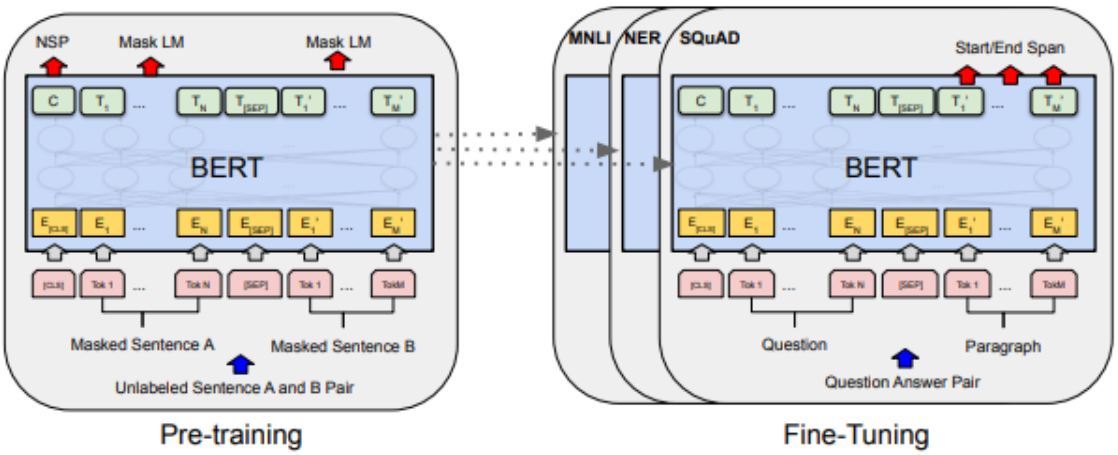


Figura 15. Proceso de Fine-Tuning sobre un modelo basado en BERT [26]

El *fine-tuning* es especialmente común en el campo de PLN y se ha vuelto muy popular con el desarrollo de modelos de lenguaje pre-entrenados de gran escala, tales como BERT, BETO o el tan nombrado GPT-3 [https://openai.com/] entre otros. El proceso de *fine-tuning* para texto generalmente sigue estos pasos:

1. **Pre-entrenamiento del modelo:** En esta etapa, se entrena el modelo de lenguaje en un corpus de texto grande y diverso. Durante el pre-entrenamiento, el modelo intenta aprender patrones y características generales del lenguaje, como gramática, sintaxis y semántica.
2. **Selección de la tarea y conjunto de datos**: El siguiente paso es decidir qué tarea específica se desea abordar para hacer el *fine-tuning* del modelo pre-entrenado. Esto podría ser, por ejemplo, clasificación de texto, etiquetado de entidades, generación de texto, etc. Luego, se necesita un conjunto de datos etiquetado específico para esa tarea en particular.
3. ***Fine-Tuning*:** Con el modelo pre-entrenado y el conjunto de datos de la tarea, se refina el entrenamiento del modelo en la tarea específica. Durante el *fine-tuning*, se ajustan los parámetros del modelo para que se adapten mejor a la nueva tarea.
4. **Evaluación y ajuste de hiperparámetros:** Una vez que se ha completado el *fine-tuning*, el modelo se evalúa en un conjunto de datos de prueba separado para medir su rendimiento en la tarea específica. Es posible ajustar los hiperparámetros del modelo para mejorar su rendimiento y generalización.

En definitiva, el *fine-tuning* es una técnica poderosa ya que permite aprovechar el conocimiento previo de modelos de lenguaje pre-entrenados, lo que a menudo resulta en un entrenamiento más rápido y una mejora en el rendimiento en tareas específicas con cantidades limitadas de datos.

5.2. Métricas de evaluación

En lo referente a las métricas de evaluación, cuatro parámetros van a ser calculados para medir el rendimiento de los modelos a entrenar. No obstante, para entender el funcionamiento de dichas métricas hace falta una definición previa.

Una matriz de confusión es una matriz que se utiliza para evaluar el rendimiento de un modelo, donde las filas representan las clases reales, las columnas representan las clases predichas por el modelo y los elementos de la matriz muestran el número de predicciones correctas e incorrectas para cada clase.

Figura 16. Matriz de confusión para problema binario

Suponiendo que el valor positivo hace referencia a los textos clasificados como ‘Productos para la piel’ y el valor negativo a ‘Productos para el cabello’:

- **Verdadero Positivo (TP):** Casos donde el valor real es positivo y el predicho por el modelo también, como por ejemplo que el Transformer clasifique un texto como ‘Productos para la piel’ y que en realidad sea así.
- **Falso Positivo (FP):** Casos donde el valor real es negativo pero el predicho por el modelo es positivo, como por ejemplo textos clasificados por el Transformer como ‘Productos para la piel’ y que pertenezcan a la clase ‘Productos para el cabello’.
- **Falso Negativo (FN):** Casos donde el valor real es positivo pero el predicho por el modelo es negativo, como por ejemplo textos clasificados por el Transformer como ‘Productos para el cabello’ y que pertenezcan a la clase ‘Productos para la piel’.
- **Verdadero Negativo (TN):** Casos donde el valor real es negativo y el predicho por el modelo también, como por ejemplo que el Transformer clasifique un texto como ‘Productos para el cabello’ y que en realidad sea así.

Una vez introducidos estos conceptos, las cuatro métricas de evaluación que se van a emplear para determinar la bondad de los modelos son:

1. ***Accuracy*:** El *accuracy* (exactitud) evalúa la proporción de predicciones correctas realizadas por un modelo en relación con el total de predicciones realizadas en un conjunto de datos.

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

2. **Precisión:** Esta métrica calcula la proporción de predicciones positivas que son verdaderamente positivas en relación con todas las predicciones positivas realizadas por el modelo.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. ***Recall*:** El *recall* mide la proporción de ejemplos positivos que son correctamente identificados por el modelo en relación con todos los ejemplos positivos en el conjunto de datos.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. ***F1-score*:** Combina precisión y *recall* en un solo valor para evaluar el rendimiento de un modelo y es especialmente útil cuando las clases están desequilibradas en el conjunto de datos.

$$F1 - score = \frac{2 \times TP}{2 \times TP + FN + FP}$$

Sin embargo, a excepción del *accuracy*, el resto de métricas sólo resuelven problemas de clasificación binaria. Es por ello que, como los textos están divididos en cinco clases, existen medidas que adaptan el resto de métricas a problemas de clasificación multiclase:

- **Macro:** Calcula la métrica por clase y luego toma el promedio no ponderado de esas métricas por clase, para dar igual importancia a todas las clases.
- **Micro:** Calcula la métrica globalmente para todas las clases combinadas y tiene en cuenta la frecuencia de cada clase, dándole más peso a las clases con mayor número de ejemplos.
- ***Weighted*:** Calcula la métrica por clase y luego toma el promedio ponderado de esas métricas por clase, utilizando el número de ejemplos de cada clase como peso.

En resumen, como se puede ver en la Tabla 5, la distribución de las clases está desbalanceada, donde más de la mitad de noticias pertenecen a la clase 'Productos para la piel', por lo que la medida más adecuada es *Weighted* para la resolución de esta tarea, ya que es útil cuando el desequilibrio de clases es significativo y se desea tener una evaluación más equitativa y representativa del rendimiento del modelo en todas las clases.

5.3. Entrenamiento de modelos

Hecha esta explicación y entendidas las métricas de evaluación, diferentes modelos de lenguaje pre-entrenados para el español van a ser analizados y comparados para ver cuál de ellos obtiene mejores resultados a la hora de hacer *fine-tuning* y entrenar ambos transformers, uno de ellos centrado en la identificación de entidades nombradas del dominio farmacéutico y el otro focalizado en la clasificación de textos farmacéuticos. Los modelos lingüísticos utilizados provienen de la biblioteca HuggingFace Transformers y son:

- **bert-base-multilingual-cased (mBERT)** [27]: Modelo pre-entrenado en más de 100 idiomas con el *dataset* más grande de Wikipedia y que enmascara ciertas palabras para intentar predecir cómo funciona y se estructura el idioma.
- **distilbert-base-multilingual-cased (DistilBERT)** [41]: Modelo basado en el anterior donde se ha llevado a cabo un proceso de 'destilación de conocimiento', una técnica de compresión en la que se entrena un modelo compacto para reproducir el comportamiento de un modelo más grande o un conjunto de modelos, con el objetivo de reducir el número de parámetros y la complejidad computacional.
- **xlm-roberta-base (RoBERTa)** [42]: XLM-RoBERTa es una versión multilingüe de RoBERTa pre-entrenado en 2,5 TB de datos filtrados de *CommonCrawl* que contienen 100 idiomas. Además, a diferencia de BERT, el enmascaramiento de las oraciones se produce en la parte de entrenamiento.
- **dccuchile/bert-base-spanish-wwm-cased (BETO)** [25]: Modelo basado en el BERT original pero entrenado con textos en castellano extraídos de diferentes organizaciones como Wikipedia, Naciones Unidas o DGT, donde se tienen en cuenta carácteres como las vocales acentuadas o la 'ñ'.

Por tanto, estos cuatro modelos de lenguaje pre-entrenados van a ser *'fine-tuneados'* con el propósito de generar un Transformer adaptado a la tarea específica a desarrollar, que será evaluado con las métricas de error comentadas anteriormente y contendrá las configuraciones de parámetros recogidas en la Tabla 8.

Learning rate	2e-5	**Weight decay**	0.01
Epochs	20	**Logging steps**	500
Train batch size & Test batch size	20	**Evaluation strategy & Save strategy**	'epoch'

Tabla 8. Parámetros de entrenamiento Transformer

Tras seleccionar los parámetros de los Transformers, se ha aplicado una técnica denominada *10-fold Cross Validation*, una técnica de evaluación de modelos en la que se divide el conjunto de datos en 10 partes iguales, se entrena y evalúa el modelo 10 veces, utilizando cada parte como conjunto de prueba una vez y el resto como conjunto de entrenamiento en cada iteración, lo que proporciona una estimación robusta del rendimiento del modelo.

Una vez hecha esta explicación, sólo queda ver los resultados obtenidos con cada uno de los modelos lingüísticos y para los dos problemas a resolver.

Resultados Transformer para NER		
mBERT	**Accuracy**	0.9696
	Precisión	0.6663
	Recall	0.6673
	F1-Score	0.6668
DistilBERT	**Accuracy**	0.9685
	Precisión	0.6402
	Recall	0.6417
	F1-Score	0.6409
RoBERTa	**Accuracy**	0.9707
	Precisión	0.6618
	Recall	0.6805
	F1-Score	0.6710
BETO	**Accuracy**	0.9704
	Precisión	0.6633
	Recall	0.6769
	F1-Score	0.6700

Tabla 9. Resultados Transformer NER

Resultados Transformer para Clasificación de textos		
mBERT	**Accuracy**	0.8258
	Weighted Precision	0.8398
	Weighted Recall	0.8258
	Weighted F1-Score	0.8025
DistilBERT	**Accuracy**	0.8108
	Weighted Precision	0.8254
	Weighted Recall	0.8108
	Weighted F1-Score	0.6760
RoBERTa	**Accuracy**	0.8258
	Weighted Precision	0.8366
	Weighted Recall	0.8258
	Weighted F1-Score	0.8113
BETO	**Accuracy**	0.8323
	Weighted Precision	0.8420
	Weighted Recall	0.8323
	Weighted F1-Score	0.8189

Tabla 10. Resultados Transformer clasificación de textos

En la Tabla 9 se observa cómo para todos los modelos, el Accuracy es bastante superior al F1-Score, mientras que en la Tabla 10 estos resultados son mucho más parecidos. Esta diferencia en la Tabla 9 referente a NER es debida a que hay una gran cantidad de etiquetas como 'O', mientras que la aparición del resto de etiquetas es muy escasa.

Por otro lado, el modelo de lenguaje pre-entrenado que mejores resultados ha dado para la tarea de NER ha sido RoBERTa, seguido muy de cerca por BETO, que será el modelo seleccionado porque su tiempo de ejecución es 1,5 veces menor al de RoBERTa, mientras que para clasificación de textos el mejor transformer ha resultado ser el basado en BETO.

A este respecto, para las dos tareas los modelos elegidos han sido los basados en BETO, por lo que ciertos parámetros van a ser variados y dichas configuraciones van a ser estudiadas para observar si es posible mejorar los resultados obtenidos.

En primer lugar, el *learning rate* controla qué tan rápido se actualizan los pesos del modelo durante el entrenamiento, así que se debe escoger un *learning rate* que sea lo suficientemente alto para converger rápidamente, pero no tan alto como para causar inestabilidad. Por tanto, el *learning rate* óptimo para NER es 2e-5, al igual que para clasificación de textos.

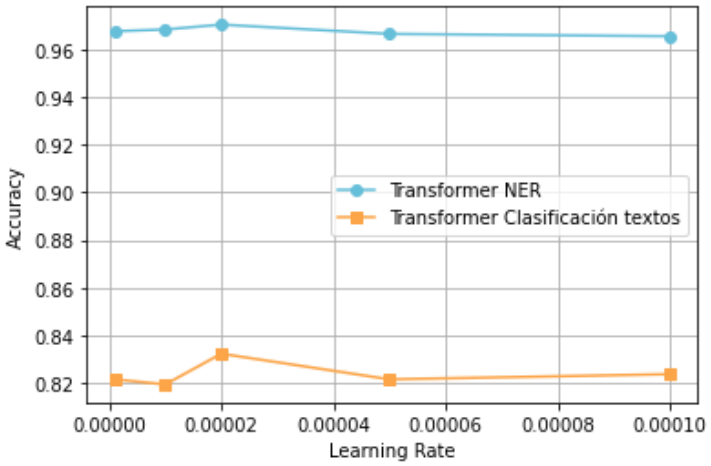


Figura 17. Valor de *learning rate* óptimo según el *Accuracy* para cada Transformer

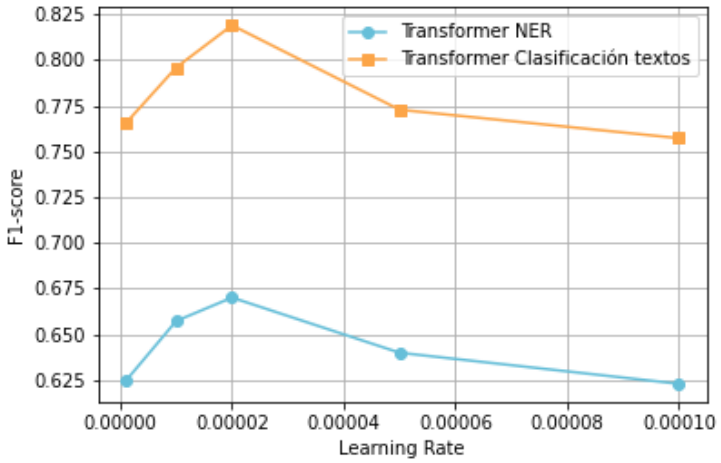


Figura 18. Valor de *learning rate* óptimo según el *F1-score* para cada Transformer

Por otro lado, un número adecuado de *epochs* (número de veces que un modelo recorre los datos durante el proceso de entrenamiento para mejorar su rendimiento) permitirá que el modelo se ajuste a los datos, pero demasiadas podrían conducir al sobreajuste. Es por ello que se ha realizado un estudio para ver qué número de *epochs* es el mejor para ambas tareas, y con las figuras situadas abajo se concluye que tanto para NER como para clasificación de textos es 20.

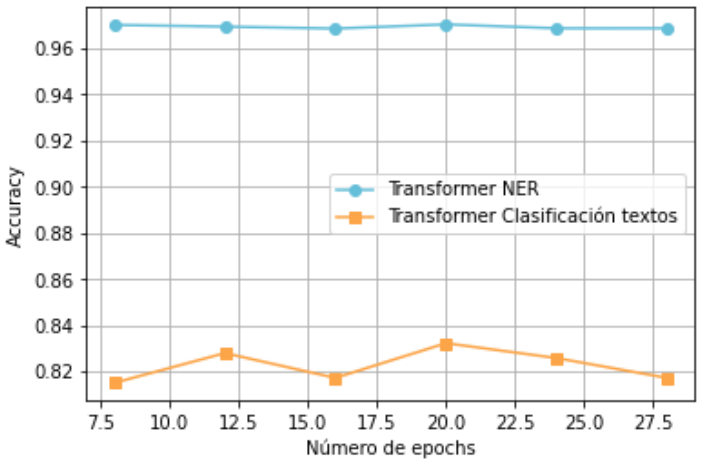


Figura 19. Número de *epochs* óptimo según el *Accuracy* para cada Transformer

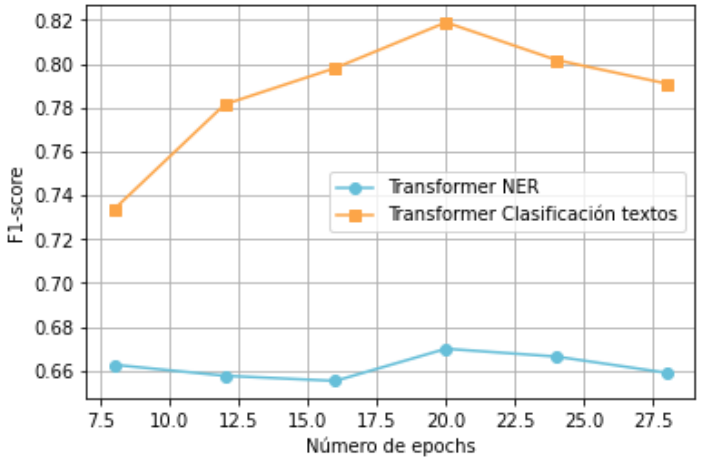


Figura 20. Número de *epochs* óptimo según el *F1-score* para cada Transformer

Finalmente, *weight decay* es una técnica que consiste en agregar una penalización a los pesos del modelo proporcional a su magnitud para evitar que los pesos tomen valores demasiado grandes y así reducir el sobreajuste. A este respecto, el valor óptimo de *weight decay* para NER es 0.01, al igual que para clasificación de textos.

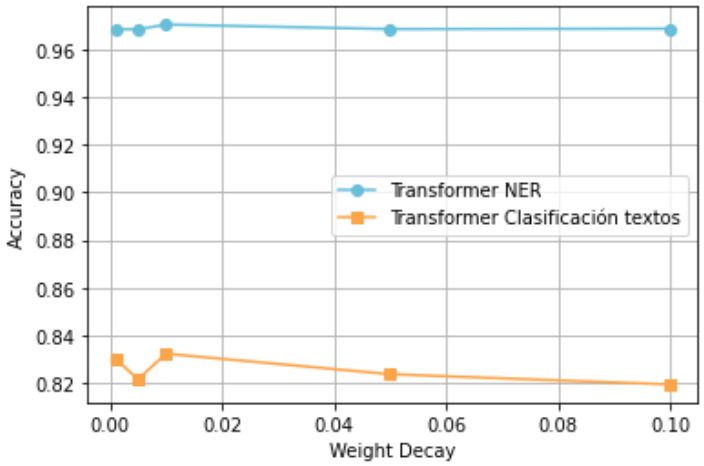


Figura 21. Valor de *weight decay* óptimo según el *Accuracy* para cada Transformer

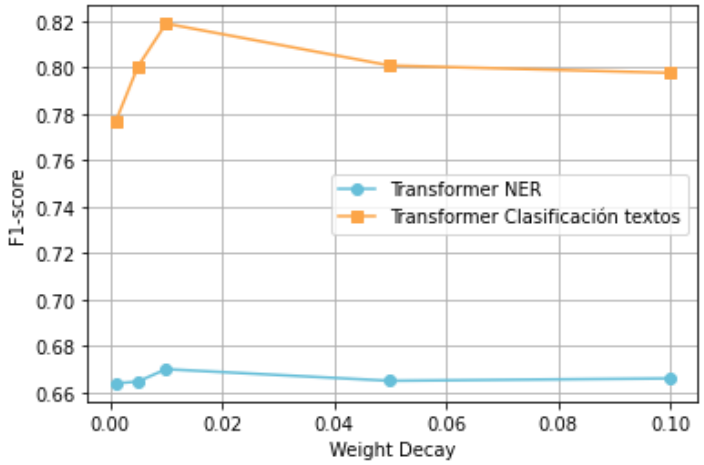


Figura 22. Valor de *weight decay* óptimo según el *F1-score* para cada Transformer

En definitiva, tras variar los valores de ciertos parámetros, se observa que los resultados óptimos coinciden con los de la configuración inicial utilizada para seleccionar el modelo de lenguaje pre-entrenado que mejor se adapta al problema y recogida en la Tabla 8.

No obstante, también es importante estudiar cómo de bien funciona el Transformer entrenado para discernir entre las cinco posibles clases referentes al ámbito farmacéutico, ya que, tal y como se ha dicho en apartados anteriores, dichas clases están desbalanceadas. Para ello en la Tabla 11 se observan las métricas calculadas para cada una de ellas, y en la Figura 23 su matriz de confusión.

	Productos para la piel	Productos para el cabello	Medicamentos / complementos	Perfumes / Desodorantes	Otros
Accuracy	0.9716	0.7288	0.7581	0.0323	0.8630
F1-Score	0.9437	0.8190	0.8356	0.0625	0.7943

Tabla 11. Resultados por clase para clasificación de textos con BETO

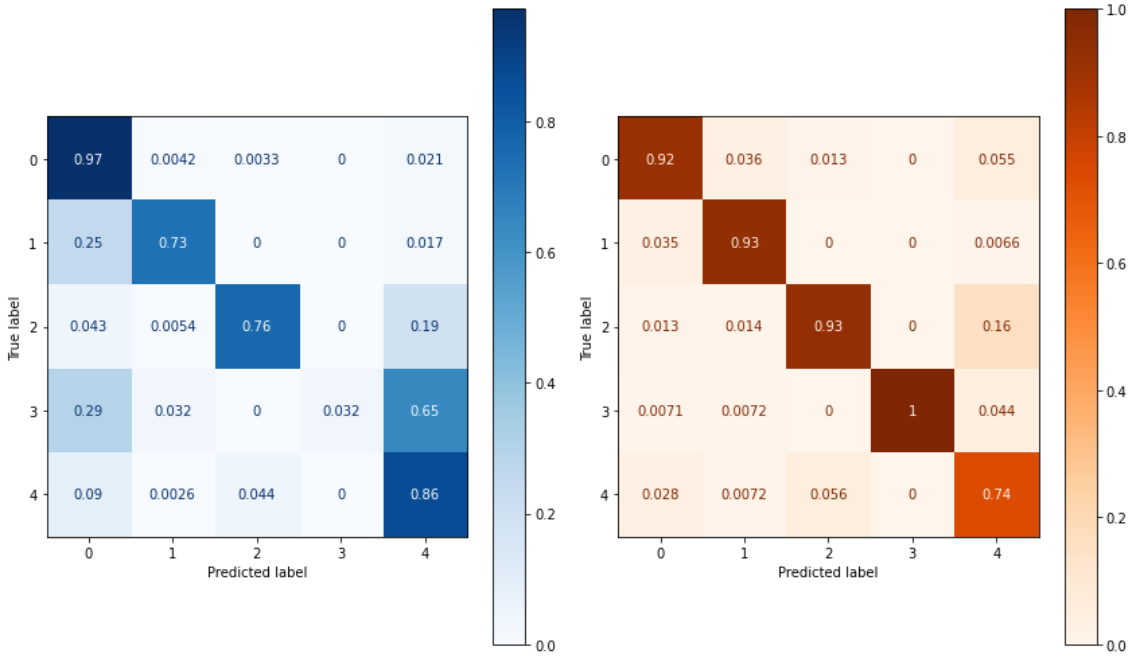


Figura 23. Matriz de confusión por clase para clasificación de textos con BETO

En resumen, las métricas obtenidas para cada una de las clases y los resultados de las matrices de confusión, donde la azul hace referencia a *recall* y la naranja a precisión, demuestran que el desbalance de las clases afecta al entrenamiento del modelo. Tanto el *accuracy* como el *F1-score* son muy altos para la primera clase 'Productos para la piel', ya que es aquella que mayor cantidad de noticias contiene, pero ambas métricas se van reduciendo a medida que el número de textos de las clases disminuye, hasta llegar a la clase 'Perfumes/desodorantes', que como se puede ver en la Tabla 5, sólo contiene 31 noticias y sus resultados son muy pobres.

Reconocimiento de entidades nombradas en el dominio farmacéutico

6. Validación y Despliegue

Una vez se han entrenado ambos Transformers para dar solución a las tareas que se tienen como objetivo, se pasa a una etapa de validación.

Para evitar el sobreajuste del modelo, se ha añadido un parámetro denominado *weight decay,* que se añade a la función de pérdida para penalizar modelos con pesos muy grandes, tal y como se ha mencionado antes. Este parámetro se ha ido modificando hasta obtener aquel que sea óptimo para cada una de las tareas a resolver.

Siguiendo este hilo, se ha optimizado de igual modo el número de *epochs* para cada Transformer, ya que un número muy alto de este parámetro también puede conducir al modelo a sobreajustar los datos.

En definitiva, se han tomado medidas para que los Transformers generados puedan funcionar bien con datos que no han visto, y a pesar de que los resultados obtenidos no son muy altos, se va a probar el funcionamiento de dichos Transformers para varios ejemplos recuperados de internet y recogidos en las cinco figuras siguientes:

Spray para aliviar de inmediato el picor, tirantez y malestar de la piel sensible o atópica.

Laboratorios Babé, marca especialista en el cuidado de la piel, lanza al mercado un innovador producto en formato spray para aliviar de forma inmediata el picor, la tirantez y el malestar en pieles sensibles o con tendencia atópica.

El nuevo Spray SOS Calmante de Babé destaca por su innovadora aplicación “touchless”, que permite aplicar el producto sin tocar la zona afectada por brotes de piel atópica, alergias o reacciones de la piel en general.

El Spray SOS Calmante de Babé, apto a partir de los 3 años, restaura la función barrera para una piel más fuerte y sana, al mismo tiempo que calma y reduce la sensación de picor y la inflamación cutánea. Un producto muy fácil de usar y aplicar gracias a su formato en spray que pulveriza una cremosa textura que, gracias a su innovadora fórmula, se gelifica al entrar en contacto con el calor de la piel. De esta manera, se evita tener que extender el producto, lo cual puede causar más molestias en una piel sensible o con brotes por dermatitis atópica.

RESULTADO:

- **Empresa:** [‘Babé’]
- **Producto:** [‘Spray SOS Calmante’]
- **Clase:** ‘Productos para la piel’

Figura 24. Prueba Transformer para la clase ‘Productos para la piel’ [43]

Sensodyne lanza pasta dental fluorada con Novamin para dientes sensibles.

El 48% de los adultos chilenos sufre de sensibilidad dental en cualquiera de sus formas, y como solución para ellos Sensodyne ofrece la primera crema dental fluorada con Novamin que puede ayudar a 'reparar y proteger' contra la hipersensibilidad de los dientes.

Aproximadamente una de cada dos personas en Chile se ven afectadas por la sensibilidad dental en algún momento de su vida. Es un dolor corto y agudo que se interpone en los placeres cotidianos como disfrutar de bebidas frías y calientes, o golosinas como el helado. Bajo esta premisa es que Laboratorio GSK desarrolló un avance en el cuidado de los dientes con la crema dental Sensodyne Repara y Protege que contiene Novamin®, una fórmula exclusiva que nace de una tecnología originariamente desarrollada para ayudar a estimular la regeneración ósea.

RESULTADO:

- **Empresa:** ['Sensodyne']
- **Producto:** ['crema dental', 'pasta dental fluorada con Novamin para dientes sensibles']
- **Clase:** 'Otros'

Figura 25. Prueba Transformer para la clase 'Otros' [44]

Normon lanza un complemento alimenticio que ayuda al funcionamiento normal del sistema inmunitario.

Normon S.A., laboratorio farmacéutico español con más de 80 años de experiencia incorpora Alyver® a su amplia gama de OTC.

La alergia es una afección muy frecuente que se presenta cuando el sistema inmune de una persona reacciona anómalamente frente a sustancias del ambiente que no son nocivas para la mayoría de la población. Como consecuencia de esa alteración del sistema inmunitario, las enfermedades alérgicas pueden producir síntomas en cualquier órgano del cuerpo, aunque son más frecuentes los problemas respiratorios, digestivos o de la piel, debido a que éstas son las zonas de mayor contacto con los agentes externos.

Alyver® es un complemento alimenticio con una innovadora fórmula patentada a base de bioactivos naturales procedentes del huevo de codorniz y el zinc, un oligoelemento que se incorpora al organismo a través de la dieta y que contribuye al funcionamiento normal del sistema inmunológico, sistema que como antes se mencionó puede verse alterado en aquellas personas sensibles a agentes externos como polen, polvo, ácaros o pelo de algunos animales.

RESULTADO:

- **Empresa:** ['Normon']
- **Producto:** ['Alyver®']
- **Clase:** 'Medicamentos/complementos'

Figura 26. Prueba Transformer para la clase 'Medicamentos/complementos' [45]

Champú en crema y sin espuma. El último lanzamiento de Carrefour.

Carrefour ha lanzado Soft champú Kéra Science, un champú en crema y sin espuma con aceite de argán y manteca de karité. Un nuevo producto dentro de su marca Les Cosmétiques, elaborado en España.

La última innovación en cosmética para cabellos de la compañía omnicanal, es un champú sin espuma y aplicado en crema que hidrata y nutre el cuero cabelludo, gracias a sus componentes ricos en aceite de argán y karité. Soft champú Kéra Science, trata el pelo a la vez que lo limpia, "desincustrando de la fibra capilar las partículas de grasa y suciedad de la forma más suave posible".

Esta nueva referencia, cuya producción se realiza en Tarragona es exclusiva de Carrefour. Además, forma parte de la gama de productos Kéra Science, dentro de su marca propia Les Cosmétiques. Esto convierte a Carrefour en "el primer distribuidor en trabajar un champú con estas características dentro de sus marcas propias". Soft champú Kéra Science ya está disponible en todos los supermercados e hipermercados de la marca, así como en la tienda online Carrefour.

RESULTADO:

- **Empresa:** ['Carrefour', 'Kéra Science']
- **Producto:** ['Soft Champú', 'Soft Champú Kéra Science']
- **Clase:** 'Productos para el cabello'

Figura 27. Prueba Transformer para la clase 'Productos para el cabello' [46]

Glamour y Cacharel celebran el lanzamiento de la fragancia 'Yes I Am Bloom Up'.

Glamour y Cacharel fueron los anfitriones de una de las fiestas de la primavera con motivo del lanzamiento de Yes I Am Bloom Up!, la nueva fragancia de Cacharel. El penthouse de WOW Concept –ubicado en la Gran Vía de Madrid– fue el lugar elegido para llevar a cabo esta celebración donde el rojo y el rosa fue el dress code sugerido para adentrarse en el universo de la icónica línea de Yes I Am.

A su llegada los invitados disfrutaron de un lugar destinado a conocer en profundidad las nuevas características del nuevo perfume. Entre los asistentes se encontraban Russian Red, Noelia López, Agoney, Flora González o Twin Melody, además de muchas otras influencers y artistas del medio nacional que disfrutaron de una gran velada gracias a las melodías de Belén Aguilera y a Pitty Bernad, quien fue la encargada de poner el broche final a la velada.

RESULTADO:

- **Empresa:** ['Cacharel', 'Glamour']
- **Producto:** []
- **Clase:** 'Otros'

Figura 28. Prueba Transformer para la clase 'Perfumes/desodorantes' [47]

Como se puede ver en las figuras, tanto el Transformer enfocado en identificar empresas y productos como aquel dedicado a clasificar textos farmacéuticos funcionan bastante bien para dar solución al problema, aunque para algunos ejemplos no consiga identificar todas las entidades de manera correcta o se equivoque al predecir una clase.

No obstante, y debido a que no se ha podido conseguir un modelo que obtenga mejores resultados, se va a proceder a explicar el despliegue que se haría en un futuro para que este modelo fuera accesible y pudiera ser ejecutado.

Para ello, en primer lugar se generará un *crawler* que, de ciertas urls que contengan información del lanzamiento de un nuevo producto, extraerá el texto y lo almacenará en una base de datos. A estos textos se les hará una limpieza y se transformarán a la estructura del etiquetado IOB tras un previo etiquetado básico.

Una vez realizado este proceso, los dos Transformers entrenados y adaptados al dominio farmacéutico se integrarán dentro de una aplicación web cuyo acceso será privado. En esta aplicación se ejecutarán las pruebas correspondientes, por las cuales el Transformer para NER recibe el texto de las noticias y devuelve la empresa y el producto lanzado por esa empresa, mientras que el Transformer para clasificación de textos predice la clase a la que pertenece dicho texto.

Esta forma de trabajar deberá ser revisada todas las semanas ya que es posible que ciertos textos no traten específicamente del lanzamiento de un nuevo producto farmacéutico o el reconocimiento de entidades o la clasificación de los textos sea errónea. Para ello se añadirá un botón de eliminación, que si se selecciona, significa que los resultados predichos por los transformers o el texto no son correctos y se procede a eliminar dicho contenido.

Cuando la revisión esté completada y se hayan borrado las noticias con errores, se entrenaran de nuevo ambos Transformers, pero añadiendo a la base de datos las nuevas noticias identificadas y clasificadas correctamente. Si alguno de los nuevos Transformers entrenados obtiene mejores resultados al ser comparado con los anteriores, será integrado en la aplicación y el modelo antiguo será desechado.

En resumen, esta es una forma de ir mejorando el funcionamiento y la clasificación de ambos Transformers a la vez que identificamos las empresas y los productos existentes en cada una de las noticias y la clase a la que pertenecen.

7. Conclusiones

A lo largo de este trabajo, se han empleado dos modelos basados en la arquitectura de Transformers, cuyo objetivo era, por un lado, identificar ciertas entidades nombradas, tales como empresa y producto, presentes en una noticia relacionada con el lanzamiento de un nuevo producto, y por otro, clasificar dichas noticias por tipo de producto.

Tras una limpieza exhaustiva de los textos recuperados de internet y la adaptación de una taxonomía para discernir entre tipos de producto, se ha realizado *fine-tuning* sobre algunos modelos pre-entrenados del lenguaje como mBERT, DistilBERT, RoBERTa o BETO, los cuales han sido usados para generar un Transformer focalizado en tareas NER y otro Transformer para clasificar los textos.

Dicho esto, el modelo que mejores resultados ha tenido para ambas tareas ha sido el basado en BETO, los parámetros del cuál se han optimizado y han dado lugar a valores de 0.97 para el *Accuracy* y de 0.67 para el *F1-Score* en el caso de Reconocimiento de Entidades Nombradas y de 0.83 de *Accuracy* y de 0.82 de *F1-score* para clasificación de textos.

No obstante, estas métricas no son muy altas debido a que la cantidad de artículos recopilados de la web no es suficiente y más de la mitad de ellos tratan de 'Productos para la piel', lo que da lugar a un desbalance de las clases.

7.1. Legado

Tal y como se ha mencionado anteriormente, el objetivo inicial de este proyecto es identificar las empresas y productos contenidos en noticias referentes al lanzamiento de un nuevo producto farmacéutico y su respectiva clasificación por tipo de producto.

Toda esta información y sobre todo la rapidez para, con un golpe de vista, conocer qué empresa está lanzando qué producto al mercado y a qué clase pertenece es sustancial para las empresas. Aquellas que tengan acceso a este contenido serán capaces de mejorar la toma de decisiones interna y llevar a cabo estrategias de negocio adaptadas a la situación actual en su nicho de mercado.

No obstante, los resultados obtenidos hasta el momento no son lo suficientemente buenos para poder realizar un despliegue y testear su funcionamiento con casos reales, pero es posible que, si se obtiene un conjunto de datos más grande y el desbalanceo entre clases se reduce, el entrenamiento y la capacidad de predicción de los modelos mejoren, y por consiguiente este proyecto pueda servir a gran cantidad de empresas para incrementar sus ventas y tomar decisiones basadas en la actualidad farmacéutica.

Además, este trabajo puede ser relevante para otras personas que tengan intereses similares de identificar ciertas entidades en un texto y clasificarlas según un criterio establecido, pero cuya temática sea diferente o el tamaño de su dataset sea mayor, para así poder obtener mejores resultados y conclusiones afines a su objetivo.

7.2. Relación del trabajo con los estudios cursados

El cimiento del presente trabajo es la creación de dos modelos basados en la arquitectura de Transformers, cuyo cometido es, por un lado, reconocer ciertas entidades nombradas que aparecen en noticias del ámbito farmacéutico, y por otro clasificar dichas noticias por tipo de producto.

Esta casuística está enormemente relacionada con los estudios cursados a lo largo del Grado en Ciencia de Datos, desde una primera etapa inicial donde se recuperan las noticias de una url en internet mediante *web scraping* y se produce una limpieza exhaustiva sobre dichos textos, hasta el entrenamiento y validación de los Transformers adaptados para dar solución a ambas tareas, pasando por un periodo de análisis exploratorio y visualización de ciertas características presentes en las noticias. Además, este análisis realizado sobre los textos fue de gran ayuda para la posterior creación de la clasificación por tipo de producto.

Por otro lado, si nos centramos en las competencias transversales adquiridas en el grado y que están relacionadas con lo expuesto en el presente trabajo, las más destacables son el Aprendizaje permanente, ya que en este proyecto he aprendido a utilizar la librería de Transformers de Hugging Face y a hacer *fine-tuning* sobre modelos del lenguaje pre-entrenados, el Análisis y resolución de problemas y la Aplicación y pensamiento práctico, a la hora de gestionar ciertos inconvenientes como la dificultad de seleccionar de forma automática las empresas y los productos en las noticias o la elaboración de un plan de trabajo y objetivos a cumplir ajustado al tiempo para redactar este documento, ítem expresamente relacionado con la competencia transversal de Planificación y gestión del tiempo.

Además, durante mi estancia en prácticas he estado trabajando con conceptos expuestos a lo largo de este documento. Uno de ellos ha sido la recuperación de noticias de internet, aunque en el caso de las prácticas eran textos relacionados con el lanzamiento de alimentos y mayoritariamente del sector cárnico.

También he profundizado en el uso de Transformers para identificar entidades relevantes o traducir textos de inglés a castellano y viceversa, así como hacer *fine-tuning* sobre modelos del lenguaje pre-entrenados para adaptarlos a una tarea específica.

Como conclusión, este trabajo ha sido de gran ayuda para reforzar conocimientos adquiridos en el grado y aprender otros nuevos, así como adquirir más destreza en ciertas competencias transversales, por lo que la relación entre los conceptos estudiados durante el grado y los expuestos en este documento es bastante estrecha.

8. Trabajos futuros

En primer lugar, un primer trabajo a realizar en un futuro sería la recopilación de más artículos relacionados con el lanzamiento de un nuevo producto de otros enlaces web, para así aumentar el tamaño de los datos a entrenar e intentar reducir el desbalance existente entre las clases de tipo de producto.

Además, gran cantidad de los anuncios dónde se habla del lanzamiento de un nuevo producto farmacéutico se publican como vídeos, audios o imágenes, para que puedan ser vistos en la televisión y en redes sociales, o escuchados en la radio por la población. Es por eso que, entrenar modelos capaces de obtener el contenido de dichos anuncios y transformarlos a texto sería otra forma de incrementar el conjunto de datos.

Por otro lado, tal y como se ha mencionado anteriormente, BERT es uno de los modelos de lenguaje pre-entrenados con mejor rendimiento y mayor éxito, por lo que se podrían utilizar otros modelos lingüísticos pre-entrenados adaptados al castellano y basados en BERT como RigoBERTa [48] o RuPERTa [49] , o incluso versiones de los Transformers ya estudiados que contengan mayor cantidad de datos de entrenamiento.

Finalmente, otra línea de trabajo futura podría ser el entrenamiento de un Transformer para que fuera capaz de identificar en el texto para que sirve el producto farmacéutico, así como que la taxonomía definida fuera más específica y detallada, es decir que, por ejemplo, para la clasificación de ‘Productos para la piel’ supiera diferenciar entre cremas anti-edad, protectores solares o geles de baño.

Reconocimiento de entidades nombradas en el dominio farmacéutico

9. Referencias

[1] Chowdhary, K.R. (2020). Natural Language Processing. En: *Fundamentals of Artificial Intelligence*. Springer, New Delhi.

[2] Translation. Hugging Face. Recuperado el 14 de junio de 2023, de https://huggingface.co/docs/transformers/tasks/translation

[3] What is Text Generation? Hugging Face. Recuperado el 14 de junio de 2023, de https://huggingface.co/tasks/text-generation

[4] Text classification. Hugging Face. Recuperado el 14 de junio de 2023, de https://huggingface.co/docs/transformers/tasks/sequence_classification

[5] Devlin J., Chang M., Lee K., & Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186*, Minneapolis, Minnesota. Association for Computational Linguistics.

[6] GPT-3. Wikipedia. Recuperado el 15 de junio de 2023, de https://es.wikipedia.org/wiki/GPT-3

[7] COVID-19: cronología de la actuación de la OMS. (2020). Recuperado el 19 de junio de 2023, de https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19

[8] Las principales librerías de Python para Web Scraping. (2023). CoDigital. Recuperado el 24 de junio de 2023, de https://codigital.ec/las-principales-librerias-de-python-para-web-scraping/

[9] Beautiful Soup vs. Scrapy vs. Selenium. (2023). Recuperado el 24 de junio de 2023, de https://www.makeuseof.com/beautiful-soup-vs-scrapy-vs-selenium

[10] Pornaras, G. (2021). Selenium vs. BeautifulSoup Python | Full Comparison. Blazemeter. Recuperado el 24 de junio de 2023, de https://www.blazemeter.com/blog/selenium-vs-beautiful-soup-python#what

[11] Valenzuela Manzanares, J. Lingüística contrastiva inglés-español: una visión general. https://cvc.cervantes.es/ensenanza/biblioteca_ele/carabela/pdf/51/51_027.pdf

[12] Diferencias estructurales entre el inglés y el español. (2013) Aprende Inglés Sila. Recuperado el 22 de junio de 2023, de https://www.aprendeinglessila.com/2013/08/diferencias-estructurales-entre-el-ingles-y-el-espanol/

[13] Baciero Fernández, José Ignacio (2020). *Elaboración de un Modelo de Reconocimiento de Entidades Nominales (NER) para su uso en aplicaciones de Procesamiento del Lenguaje Natural (NLP)*. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. Industriales (UPM), Madrid, España.

[14] Rodríguez, F. (2023). ¿Cómo implementar un NER? Keep Coding. Recuperado el 13 de junio de 2023, de https://keepcoding.io/blog/como-implementar-un-ner/

[15] S. Hochreiter y J. Schmidhuber. (1997) Long Short-term Memory. En *Neural computation*, *vol. 9, págs. 1735-1780*.

[16] Sarkar, S. (2021). Named Entity Recognition using Deep Learning(ELMo Embedding+ Bi-LSTM). Medium. Recuperado el 13 de junio de 2023, de https://medium.com/analytics-vidhya/named-entity-recognition-using-deep-learning-elmo-embedding-bi-lstm-48295bc66cab

[17] Text Classification: What it is And Why it Matters. MonkeyLearn. Recuperado el 14 de junio de 2023, de https://monkeylearn.com/text-classification/

[18] S. Ranganathan, K. Nakai, & C. Schonbach. (2019). Bayes' Theorem and Naive Bayes Classifier. En *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, págs. 403-412*.

[19] Ming Leung, K. (2007). Naive Bayesian Classifier. En *Computer Science and Engineering*.

[20] Noble, W. (2006). What is a support vector machine?. En *Nat Biotechnol 24, págs. 1565–1567*.

[21] Brownlee, J. (2020). One-vs-Rest and One-vs-One for Multi-Class Classification. Machine Learning Mastery. Recuperado el 29 de junio de 2023, de https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

[22] ¿Qué son las redes neuronales recurrentes? IBM. Recuperado el 30 de junio de 2023, de https://www.ibm.com/es-es/topics/recurrent-neural-networks

[23] Pengfei, L., Xipeng, Q., & Xuanjing, H. (2023). Recurrent Neural Network for Text Classification with Multi-Task Learning. *En Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, págs. 2873-2879.*

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. En *Advances in neural information processing systems*.

[25] Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., & Pérez, J. (2020) Spanish Pre-Trained BERT Model and Evaluation Data. En *PML4DC at ICLR 2020.*

[26] Devlin J., Chang M., Lee K., & Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186*, Minneapolis, Minnesota. Association for Computational Linguistics.

[27] Devlin, J., Chang, W., Lee, K., & Toutanova, K. (2023). Bert-base-multilingual-cased · Hugging Face. Hugging Face. Recuperado el 3 de julio de 2023, de https://huggingface.co/bert-base-multilingual-cased

[28] Fine-tuning a un modelo pre-entrenado. Hugging Face. Recuperado el 3 de julio de 2023, de https://huggingface.co/docs/transformers/v4.19.0/es/training

[29] Política de Privacidad | Podium GM. Podium Global Media. Recuperado el 11 de julio de 2023, de https://www.podiumgm.com/index.php/avisolegal/

[30] Lanzamientos. Farmaventas. Recuperado el 23 de junio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos.html

[31] Davlan/distilbert-base-multilingual-cased-ner-hrl. Hugging Face. Recuperado el 7 de julio de 2023, de https://huggingface.co/Davlan/distilbert-base-multilingual-cased-ner-hrl

[32] Productos Farmacéuticos: Clasificación y Características. (2018). Esneca Business School. Recuperado el 5 de julio de 2023, de https://www.esneca.com/blog/tipos-de-productos-farmaceuticos/

[33] Text classification. Hugging Face. Recuperada el 7 de julio de 2023, de https://huggingface.co/docs/transformers/v4.17.0/en/tasks/sequence_classification

[34] Recurrent neural network unfold.svg. Wikimedia Commons. Recuperado el 30 de junio de 2023, de https://commons.wikimedia.org/wiki/File:Recurrent_neural_network_unfold.svg

[35] Cantabria Labs presenta Neoretin Discrom Control Pigment Neutralizer Serum. (2023). Farmaventas. Recuperado el 10 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/13948-cantabria-labs-presenta-neoretin-discrom-control-pigment-neutralizer-serum.html

[36] Nuevo Vinoperfect Tratamiento Ojos Iluminador de Caudalie. (2023). Farmaventas. Recuperado el 10 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/13951-nuevo-vinoperfect-tratamiento-ojos-iluminador-de-caudalie.html

[37] Agua de mar para prevenir las alergias otoñales. (2015). Farmaventas. Recuperado el 12 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/4229-agua-de-mar-para-prevener-las-alergias-otonales.html

[38] Ferrer presenta una nueva combinación de Omega-3, OM3GAFORT. (2023). Farmaventas. Recuperado el 12 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/4838-ferrer-presenta-una-nueva-combinacion-de-omega-3-om3gafort

[39] Nace MartiDerm SUN CARE, la nueva gama de fotoprotectores. (2023). Farmaventas. Recuperado el 12 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/13920-nace-martiderm-sun-care-la-nueva-gama-de-fotoprotectores.html

[40] Babé presenta Super Fluid Depigment+, el protector solar despigmentante que protege, previene y atenúa las manchas. (2023). Farmaventas. Recuperado el 14 de julio de 2023, de https://www.farmaventas.es/ultimos-lanzamientos/13946-babe-presenta-super-fluid-depigment-el-protector-solar-despigmentante-que-protege-previene-y-atenua-las-manchas-del-rostro.html

[41] Sanh, V. (2023). distilbert-base-multilingual-cased · Hugging Face. Hugging Face. Recuperado el 19 de julio de 2023, de https://huggingface.co/distilbert-base-multilingual-cased

[42] xlm-roberta-base · Hugging Face. Hugging Face. Recuperado el 19 de julio de 2023, de https://huggingface.co/xlm-roberta-base

[43] Spray para aliviar de inmediato el picor, tirantez y malestar de la piel sensible o atópica. (2022). Nutrasalud. Recuperado el 21 de agosto de 2023, de https://www.nutrasalud.es/productos/20220201/laboratorios-babe-spray-sos-calmante-piel-atopica-sensible

[44] Sensodyne lanza pasta dental fluorada con Novamin para dientes sensibles. (2015). Mujeres y más. Recuperado el 21 de agosto de 2023, de https://mujeresymas.cl/sensodyne-lanza-pasta-dental-fluorada-con-novamin-para-dientes-sensibles/

[45] Normon lanza un complemento alimenticio que ayuda al funcionamiento normal del sistema inmunitario. Laboratorios Normon España. Recuperado el 21 de agosto de 2023, de https://www.normon.es/noticia/normon-lanza-un-complemento-alimenticio-que-ayuda-al-funcionamiento-normal-del-sistema-inmunitario

[46] Davara, A. (2020). Soft champú Kéra Science, el último lanzamiento de Carrefour. Distribución / Actualidad. Recuperado el 21 de agosto de 2023, de https://www.distribucionactualidad.com/soft-champu-kera-science-lanzamiento-carrefour/

[47] Palacios, R., & Belda, Á. (2023). Glamour y Cacharel celebran el lanzamiento de la fragancia 'Yes I Am Bloom Up'. Vogue España. Recuperado el 21 de agosto de 2023, de https://www.vogue.es/articulos/glamour-cacharel-lanzamiento-perfume-yes-i-am-bloom-up-evento

[48] Vaca A., Garcia G., Montoro H., Aldama N., Doaa S., Betancur D., Moreno A., Guerrero M. & Barbero A. (2022). RigoBERTa: A State-of-the-Art Language Model For Spanish. Instituto de ingeniería del conocimiento. Universidad Autónoma de Madrid.

[49] mrm8488/RuPERTa-base: Spanish RoBERTa. GitHub. Recuperado el 31 de agosto de 2023, de https://github.com/mrm8488/RuPERTa-base

[50] Briceño, B., & Fernández, E. (2021). ¿Qué son los word embeddings y para qué sirven? Blogs iadb. Recuperado el 31 de agosto de 2023, de https://blogs.iadb.org/conocimiento-abierto/es/que-son-los-word-embeddings/

Anexos

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**				-
ODS 2. **Hambre cero.**				-
ODS 3. **Salud y bienestar.**	-			
ODS 4. **Educación de calidad.**				-
ODS 5. **Igualdad de género.**				-
ODS 6. **Agua limpia y saneamiento.**				-
ODS 7. **Energía asequible y no contaminante.**				-
ODS 8. **Trabajo decente y crecimiento económico.**	-			
ODS 9. **Industria, innovación e infraestructuras.**		-		
ODS 10. **Reducción de las desigualdades**.				-
ODS 11. **Ciudades y comunidades sostenibles**.				-
ODS 12. **Producción y consumo responsables.**		-		
ODS 13. **Acción por el clima.**				-
ODS 14. **Vida submarina.**				-
ODS 15. **Vida de ecosistemas terrestres.**				-
ODS 16. **Paz, justicia e instituciones sólidas.**				-
ODS 17. **Alianzas para lograr objetivos.**			-	

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Un modelo Transformer dedicado a identificar productos y empresas en el contexto de lanzamiento de nuevos productos farmacéuticos puede tener un impacto positivo en la consecución de varios Objetivos de Desarrollo Sostenible, contribuyendo a la mejora de la salud y el bienestar de las personas, la innovación en el sector de la salud y la promoción de prácticas responsables y sostenibles en la industria farmacéutica. A este respecto, los ODS más relacionados con el presente trabajo son:

- **ODS 3: Salud y bienestar:** La identificación precisa de nuevos productos farmacéuticos y sus empresas puede contribuir a mejorar la salud y el bienestar de la población al identificar nuevas y prometedoras soluciones farmacéuticas y terapias innovadoras, así como al facilitar el acceso, tanto a información relevante sobre tratamientos y terapias, como a medicamentos seguros y efectivos que contribuyan a una vida sana. Además, esta identificación de entidades farmacéuticas puede ayudar a los profesionales de la salud, gobiernos y pacientes a tomar decisiones informadas sobre tratamientos y opciones terapéuticas.
- **ODS 8: Trabajo decente y crecimiento económico:** El desarrollo de tecnologías avanzadas como los Transformers, utilizados para identificar nuevos productos farmacéuticos y sus respectivas empresas, puede impulsar la creación de nuevos medicamentos y tratamientos, que a su vez pueden contribuir a mejorar el crecimiento económico, a aumentar la productividad de las empresas farmacéuticas y a fomentar la inversión en nuevas tecnologías relacionadas con el sector.
- **ODS 9: Industria, innovación e infraestructuras:** La utilización de Transformers para identificar productos y empresas en el sector farmacéutico representa una aplicación de tecnología innovadora en la industria de la salud, lo que fomenta el desarrollo y la mejora de infraestructuras y procesos relacionados con la investigación y desarrollo de medicamentos, así como la creación de nuevas líneas de trabajo basadas en la información obtenida de los productos con más ventas y más demandados en el nicho de mercado farmacéutico.
- **ODS 12: Producción y consumo responsables:** La identificación de empresas farmacéuticas y sus productos puede contribuir a un enfoque más responsable y sostenible en la producción y consumo de medicamentos, permitiendo una mejor supervisión de las prácticas de las empresas y la calidad de los productos farmacéuticos.
- **ODS 17: Alianzas para lograr objetivos:** La aplicación de modelos de lenguaje avanzados como los Transformers en el campo de la farmacéutica puede fomentar alianzas y colaboraciones entre el sector público y privado, así como con organizaciones internacionales y académicas, para promover la investigación y el desarrollo de medicamentos más efectivos y asequibles.