UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Anonimización Personalizada

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Durá Santonja, Carlos

Tutor/a: Monserrat Aranda, Carlos

Cotutor/a externo: GARCIA MARTINEZ, MARIA MERCEDES

CURSO ACADÉMICO: 2021/2022

Resum

El desenvolupament i l'expansió de les noves tecnologies plantegen un nou repte, el de continuar amb el progrés tecnològic alhora que es garanteix la privacitatdels seus usuaris. Aquí és on entra en joc l'anonimització; aquest procés serveix per protegir les dades personals sensibles mitjançant diferents tècniques. L'objectiu d'aquest treball final de grau és el desenvolupament d'un model capaç d'anonimitzar text no estructurat adaptant-se a les necessitats de l'usuari.

Per desenvolupar aquest treball s'han utilitzat models de reconeixement d'entitats pre-entrenats de *spaCy*. Per al *fine tunning* dels models s'han utilitzat principalment les dades pre-anotades del projecte wikiner. A banda dels reentrenaments, també s'ha afegit al model la possibilitat de personalitzar-lo a través de diferents opcions com l'anonimització via expressions regulars o l'anonimització forçada a través de llistes d'entitats, entre d'altres opcions.

Finalment, s'ha realitzat una anàlisi dels resultats obtinguts avaluant la correcta anonimització de diferents textos i les millores respecte al model base de *spaCy*.

Paraules clau: Traducció Automàtica, Estimació de qualitat de la traducció automàtica, Europeana Translate, Aprenentatge Automàtic, Aprenentatge Profund, Xarxes Neuronals

Resumen

El desarrollo y la expansión de las nuevas tecnologías plantean un nuevo reto, el de continuar con el progreso tecnológico al tiempo que se garantiza la privacidad de sus usuarios. Aquí es donde entra en juego la anonimización. Este proceso sirve para proteger los datos personales sensibles mediante diferentes técnicas. El objetivo de este trabajo final de grado es el desarrollo de un modelo capaz de anonimizar texto no estructurado adaptándose a las necesidades de un usuario.

Para el desarrollo de este trabajo se han utilizado modelos de reconocimiento de entidades pre-entrenados de *spaCy*. Para el *fine tunning* de los modelos se han utilizado principalmente los datos anotados del proyecto wikiner, incluyendo anotaciones de los mismos datos de nuevas etiquetas, como profesión o nacionalidad. A parte de los reentrenamientos también se ha añadido al modelo la posibilidad de personalizarlo a través de diferentes opciones como la anonimización vía expresiones regulares o la anonimización forzada a través de listas de entidades, entre otras opciones.

Por último, se ha realizado un análisis de los resultados obtenidos evaluando la correcta anonimización de diferentes textos y las mejoras respecto al modelo base de *spaCy*.

Palabras clave: Traducción Automática, Estimación de calidad de la traducción automática, Europeana Translate, Aprendizaje Automático, Aprendizaje Profundo, Redes Neuronales

Abstract

The development and expansion of new technologies pose a new challenge, that of continuing technological progress while ensuring the privacy of its users. This is where anonymization comes into play, this process serves to protect sensitive personal data using different techniques. The objective of this final degree work is the development of a model capable of anonymizing unstructured text adapting to the needs of a user.

For the development of this work, pre-trained entity recognition models of *spaCy* have been used. For the *fine tuning* of the models we mainly used the annotated data from the wikiner project. Apart from the re-training, we have also added to the model the possibility of customizing it through different options such as anonymization via regular expressions or forced anonymization through lists of entities, among other options.

Finally, an analysis of the results obtained has been carried out, evaluating the correct anonymization of different texts and the improvements with respect to the *spaCy* base model.

Key words: Machine Translation, Machine Translation Quality Estimation, Europeana Translate, Machine Learning, Deep Learning, Neural Networks

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Este capítulo introductorio brinda al lector un contexto general para enmarcar el trabajo presentado en este documento. A continuación, describimos la motivación que nos llevó a realizar este trabajo y los objetivos propuestos del mismo. Además, se realiza una breve ontroducción al campo de la *Anonimización*.

Por último, se resume la estructura del resto del presente TFG, permitiendo al lector tener una visión general del trabajo realizado.

1.1 Motivación

Con la evolución de nuevas tecnologías y la digitalización de todo tipo de documentos como contratos, expedientes o facturas ha surgido la necesidad de la protección de los datos que estos contienen. Las medidas que ha adoptado el *Parlamento Europeo* se recogen en el *Reglamento General de Protección de Datos* (RGPD), creando de esta forma una normativa a seguir para el correcto tratamiento de los datos personales y su libre circulación. Por ello, en el campo del procesamiento del lenguaje natural o en inglés *Natural Language Processing* (NLP) ha crecido el interés en crear modelos de reconocimiento de entidades o en inglés *Named Entity Recognition* (NER) capaces de detectar estos datos sensibles y ocultarlos para una correcta utilización de los documentos en los cuales aparecen. El reconocimiento de entidades, o lo que deriva de ello, la anonimización de texto representa un desafío para el campo del procesamiento del lenguaje, debido a las distintas características que pueden adoptar los diferentes idiomas.

1.2 Objetivos

El objetivo principal de este proyecto consiste en el estudio y la experimentación de diferentes técnicas de anonimización en texto no estructurado. El trabajo se divide en tres objetivos:

- Creación de una herramienta de anonimización: desarrollar una herramienta capaz de anonimizar texto no estructrado a través de diferentes técnicas.
- Reentrenamiento de modelos NER: hacer el *Fine tunning* de los modelos base de spaCy para mejorar el reconocimiento de entidades de forma automática y así mejorar la anonimización.
- Capacidad de personalización: dotar al modelo de anonimización final de la capacidad para personalizar la anonimización del texto según las necesidades.

1.3 Anonimización

La anonimización es un problema reciente en la industria del procesamiento del lenguaje natural o en inglés *Natural Processing Language* (NLP) y las manipulaciones de bases de datos. La resolución de este problema consiste en el enmascaramiento o eliminación de datos sensibles que pueda contener un documento o una base de datos.

En el mundo que estamos viviendo hoy en día, un mundo muy globalizado, donde constantemente se envian documentos privados o oficiales con información sensible o se almacenan estos datos a través de formularios, registros, o en webs, es de vital importancia tener todos los datos y documentos controlados y protegidos, para que de esta forma no corran ningún peligro los datos privados de cualquier persona física o entidad. Todo ello supone un gran reto a la hora de manipular esos textos o datos para una correcta utilización y protección de estos mismos.

Los sistemas de anonimización automática están basados en sistemas de reconocimiento de entidades nombradas. Los primeros sistemas que se utilizaron fueron sistemas basados en reglas, capaces de identificar determinados tipos de entidades con una estructura muy marcada. Estos sistemas de reconocimiento de entidades son creados a partir de expertos de forma manual. Aunque estas reglas quedan bien definidas y con una estructura sólida, el proceso es costoso y muchas reglas no son capaces de recoger muchos dominios de vital importancia a la hora de la anonimización como pueden ser nombres o ciudades.

Con el paso del tiempo, los sistemas basados en reglas se fueron sustituyendo por los modelos de aprendizaje automático supervisado. Estos modelos permitían detectar entidades sin necesidad de tener que establecer previamente de forma manual reglas escritas. Por ende, este sistema era mucho más rápido y menos costoso que el basado en reglas. Pero en estos últimos años estos modelos se han visto superados por técnicas de aprendizaje profundo o Deep Learning en inglés, ya que estos modelos eran capaces de recoger el contexto de una mejor forma para la detección nuevos tipos de entidades.

1.4 Estructura de la memoria

El presente trabajo está constituido por un total de cinco capítulos. A continuación, se describen los contenidos tratados en cada capítulo:

En el **capítulo 1**, se realiza una introducción del tema que se aborda a lo largo del documento, además se detalla la motivación y los objetivos del trabajo fin de grado.

En el **capítulo 2**, se expone de forma teórica las diferentes aproximaciones de la anonimización automática existentes y sus características principales. Además, se presentan las distintas técnicas empleadas en NLP.

En el **capítulo 3**, se presenta al lector las distintas técnicas de anonimización automática que se han empleado en el presente proyecto para la anonimización de textos, las características particulares de cada método y las posibles personalizaciones que se pueden aplicar a la anonimización.

En el **capítulo 4**, se muestra al lector el marco experimental del proyecto, el cual engloba las tecnologías que han sido empleadas en el desarrollo del trabajo, la forma de evaluación de los modelos de anonimización entrenados, la preparación de los datos iniciales, la creación de los distintos conjuntos de datos generados, los experimentos que se han realizado a lo largo del proyecto.

En el **capítulo 5**, se presentan los resultados obtenidos a partir de la experimentación realizada y de los modelos utilizados a lo largo del proyecto

En el **capítulo 6**, se presentan las conclusiones obtenidas a partir de la experimentación realizada y se exponen diferentes caminos a tomar en próximos trabajos relacionados con la anonimización automática.

CAPÍTULO 2
Estado del arte

En el presente capítulo, se presenta al lector la historia de las tecnologías más relevantes en el campo de la anonimización automática y las diferentes técnicas empleadas.

2.1 Reglamento General de Protección de Datos

El *Reglamento general de Protección de Datos, RGPD*, es el reglamento europeo encargado de la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de los mismos. Entró en vigor en 2016, y fue de aplicación en 2018, dejando de esta forma a las empresas y organizaciones dos años para que se adaptasen a este nuevo reglamento. Es una normativa a nivel de la Unión Europea, de forma que cualquier empresa que está dentro de la unión y maneje información personal debe de cumplimentar el reglamento.

El nuevo reglamento de protección de datos de la Unión Europea amplía el alcance de la ley de protección de datos de la unión a todas las compañías extranjeras que tratan datos de residentes de la UE. El RGPD también ofrece un conjunto de "derechos digitales"para los ciudadanos de la Unión Europea en una época en la que el valor de los datos personales está aumentando cada vez más.

En España, la *Ley orgánica de Protección de Datos de Carácter Personal, LOPD*, se vió obsoleta por el RGPD, siendo sustituida por la *Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales, LOPD-GDD*, acorde con el reglamento Europeo.

2.2 Modelos de Reconocimiento de Entidades Nombradas

Los modelos de reconocimiento de entidades (NER) tratan de extraer la información del texto y clasificarla en categorías predefinidas mediante diferentes estructuras de una red neuronal de *Deep Learning*. Dentro de los diferentes software que disponen de NER se pueden destacar tres.

2.2.1. SpaCy

SpaCy [8] es una librería de Python que permite construir aplicaciones del procesamiento del lenguaje natural (NLP). SpaCy proporciona modelos preentrenados para diferentes idiomas y de diferentes volúmenes, dependiendo de si quieres un modelo más lento pero preciso y que ocupa mucho espacio, o modelos de menor volumen que funcionan más rápido pero con menor precisión.

Dentro de las diferentes herramientas que dispone spaCy en la pipeline de cada modelo, podemos encontrar tokenizer, tagger, ner, parser, entre otras.

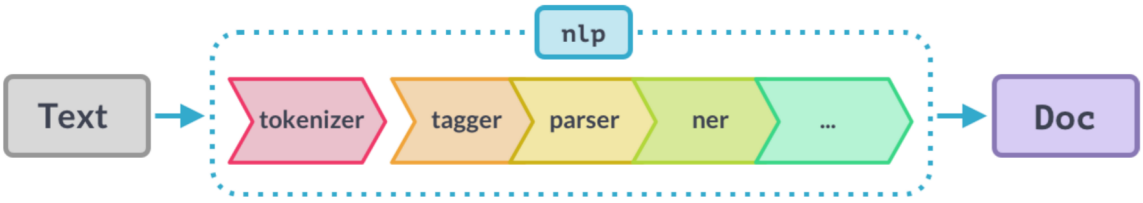

Figura 2.1: *Pipeline de SpaCy.*

Además de los modelos preentrenados, spaCy tiene la capacidad de crear nuevos modelos o reentrenar los modelos preentrenados para dotar al modelo resultante de una mayor personalización.

2.2.2. NLTK

NLTK [3], *Natural Language Toolkit* es un conjunto de librerías juntadas para su uso en el mundo del procesamiento del lenguage natural. Proporciona algoritmos sencillos para el análisis de todo tipo recursos léxicos junto con un conjunto de bibliotecas de procesamiento de texto. Se utiliza ampliamente en la investigación y con fines educativos.

Está diseñado para trabajar con una amplia gama de disciplinas, incluyendo el modelado de investigación o la inteligencia artificial. También es adecuado para expertos de la industria del *bootstrapping*.

La librería proporciona frameworks para más de 50 corpus y recursos léxicos. Soporta razonamiento semántico, clasificación, análisis sintáctico, derivación y etiquetado permitiendo programar funciones para el NLP.

NLTK realiza el NER en dos pasos. El primer paso es el POS-tagging o el etiquetado gramatical, al que sigue la fragmentación (o chunking en inglés) para extraer las entidades nombradas.

2.2.3. Flair

Flair [1] es un framework basado en Python para el procesamiento de lenguaje natural (NLP). Permite a los usuarios realizar tareas estándar de NLP, como reconocimiento de entidades nombradas (NER), part-of-speech tagging (PoS), desambiguación y clasificación del sentido de las palabras, funciona bien en diferentes tareas del procesamiento del lenguaje.

Flair presenta una interfaz sencilla y unificada para una variedad de *embeddings* [7] de documentos y palabras, incluyendo BERT, Elmo y las propias de Flair. También tiene soporte multilingüe. El framework en si está construido sobre PyTorch, uno de los frameworks más potentes y utilizados a día de hoy en el campo del *Deep Learning*.

A parte de los modelos entrenados para las diferentes tareas de NLP, tiene la opción de crear nuevos modelos customizados por el usuario. Esta librería es capaz de soportar más de 250 idiomas, por lo que es útil para crear modelos pequeños.

2.3 Anonimización en NLP

El interés sobre la anonimización en texto plano ha crecido durante los últimos años debido al desarrollo de los sistemas de protección de datos. Esta tarea es tediosa y monótona para una persona por lo que los desarrolladores han comenzado a crear sistemas automáticos que consigan una mejor y más eficiente protección de los datos.

Para conseguir desarrollar estos sistemas se ha necesitado detectar entidades personales dentro de un mismo texto mediante el uso de técnicas NLP. Las posibles entidades podrían ser el nombre de una persona, su DNI, una ciudad o una organización entre otras posibles entidades.

Para el reconocimiento de estas entidades dentro de documentos de texto, se han utilizado diferentes métodos como reglas de expresión regular para detectar entidades con un formato exacto, diccionarios donde están las entidades que se quieren anonimizar o modelos de redes neuronales entrenados para reconocer diferentes tipos de entidades en un texto.

2.3.1. Anonimización por Reglas

La primera aproximación que se hizo sobre el reconocimiento de entidades nombradas para su posterior anonimización fue en 1991 en la conferencia en la que se presentó la propuesta [15].

Esta propuesta se orientó únicamente al reconocimiento de compañías basándose en heurísticas y en reglas escritas. Estas reglas de expresión regular son muy utilizadas para la búsqueda de cadenas de texto con un formato exacto (como podría ser el de un DNI con 8 números y una letra) en un documento

2.3.2. Anonimización por Diccionarios

Derivada de la primera aproximación al reconocimiento de entidades surgieron dos ramas: una de estas ramas continuó la propuesta inicial centrándose en las expresiones regulares y diccionarios como se pudo ver en la propuesta [17].

Esta propuesta detectaba 200 entidades en una combinación de expresiones regulares y diccionarios. Estos diccionarios se dividen en un diccionario por cada tipo de entidad, conteniendo cada diccionario el nombre de las diferentes entidades que pertenecen a su tipo.

2.3.3. Anonimización Automática

La otra rama que surgió a partir de la primera propuesta fue la de aplicar y crear técnicas y modelos de Aprendizaje Automático en textos. Las primeras propuestas en esta rama utilizaron modelos basados en Árboles Binarios [16] y Máquinas de Vectores Soporte [2], con esta implementación se pasó de un único tipo de entidad reconocida en 1991 a 8 diferentes tipos de entidad con hasta un 87 % de F1.

Aunque hubo una gran mejora con estos modelos, debido al gran crecimiento del Deep Learning en los últimos años, surgió la posibilidad de realizar experimentos con redes neuronales con diferentes tipos de arquitectura entre las que cabe destacar las Redes Neuronales Convolucionales, las Redes Neuronales Recurrentes [4] y los Transformers [20]. Estos diferentes modelos dieron un gran salto en el reconocimiento de entidades nombradas y,como consecuencia, en la anonimización de dichas entidades.

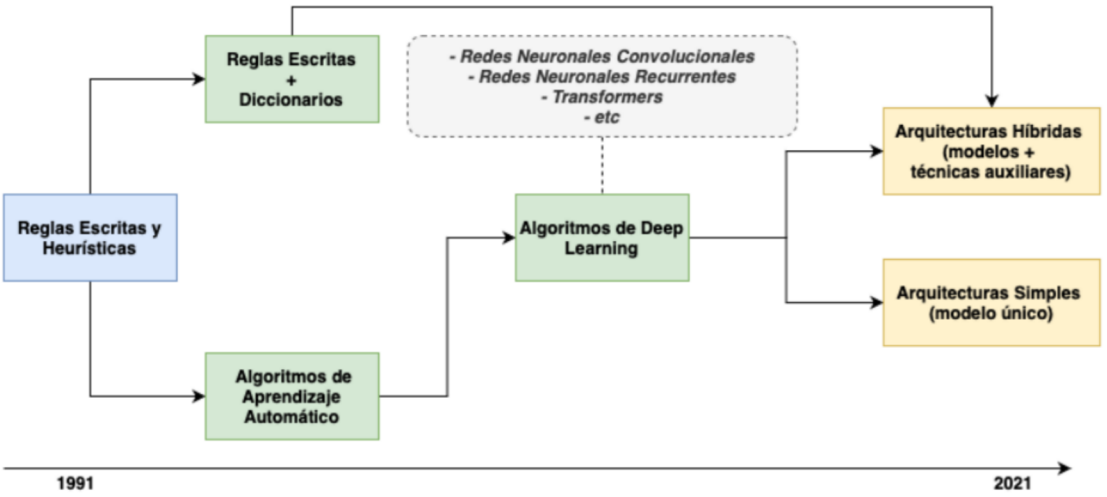

Figura 2.2: Evolución tecnológica en procesos NER [18].

2.4 Anonimización en Bases de Datos

2.4.1. K-Anonimización

La K-anonimización [13] es una técnica que permite cuantificar y aplicar un grado de anonimización a la información de los sujetos presentes en una base de datos. Para ello aplica diferentes técnicas a los diferentes tipos de datos que son:

- Los **datos identificadores** son aquellos que son capaces de identificar por si solos a un sujeto de forma inequívoca, como puede ser el nombre de una persona, el DNI o el número de teléfono. La K-anonimización no trabaja con este tipo de datos, ya que al ser capaces de identificar a un sujeto por si solos, los elimina directamente.
- Los datos **cuasi-identificadores** son aquellos datos que por sí solos no son capaces de identificar a una persona, pero que en conjunto con otros datos, sí que puede llegar a identificarlo, como sería por ejemplo, el código postal, la fecha de nacimiento o el género. La K-anonimización sí que trabaja con estos datos, de forma que ya no puedan llegar a identificar en conjunto.
- Los **datos sensibles** son aquellos que pueden resultar muy comprometidos para la privacidad de una persona, como podría ser una enfermedad. La K-anonimización busca que estos datos no puedan ser relacionados con los cuasi-identificadores.

La K-anonimización mide las probabilidades de que un tercero pueda relacionar los datos que hayan sido tratados, y la posibilidad de llegar a una persona a la que se distribuyen dichos datos.

Un individuo es k-anónimo, dentro de la base de datos que se encuentra, cuando, y solo cuando, para cualquier combinación de atributos cuasi-identificadores existan al menos K individuos con esos mismos valores en los atributos.

De esta forma la probabilidad de identificar a una persona es como máximo $1/K$, por lo que para garantizar un bajo riesgo de posible identificación del individuo hace falta establecer un valor mínimo alto de K cuando se desee llevar a cabo un proceso de anonimización. Para llevar la K-anonimización a cabo se pueden realizar dos métodos distintos.

- **Generalización**: la generalización consiste en transformar o generalizar los valores de los elementos cuasi-identificadores dentro de un conjunto de valores o un intervalo de tal forma que estos sean menos precisos. Esto se puede hacer mediante la creación de rangos en el caso de valores numéricos. Por ejemplo, si la edad exacta de un usuario está entre 20 y 30 sustituir la edad por el intervalo 20-30, o mediante la creación de jerarquías en caso de valores nominales, como en el atributo *Deportes*, sustituir *Baloncesto* o *Fútbol*, englobarlo en *Deportes con balón*.
- **Eliminación**: este método es utilizado en el caso de encontrar algún registro que contenga un valor que no se puede generalizar de forma que se siga conservando suficiente información para un análisis de los datos. Si se encuentra con este caso se opta por como bien dice el nombre del método, la eliminación de este registro, de modo que este no contamine el conjunto de datos y distorsione los resultados. También hay que eliminar los registros cuyos valores son muy poco usuales, es decir, que contenga datos atípicos.

2.5 Pseudonimización

La pseudonimización de datos personales es según el RGPD[1] "*aquella información que, sin incluir los datos denominativos de un sujeto, permiten identificarlo mediante información adicional, siempre que esta figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable*".

En otras palabras, la pseudonimización [6] consiste en tratar los datos personales sin los datos identificativos, pero sin suprimir la vinculación entre los datos que identifican a la persona a la cual pertenecen los datos, guardando esta vinculación en un documento a parte, con el cual se pueda realizar la reversión.

Para la correcta pseudonimización es muy importante la custodia de información adicional que permite vincular el dato pseudonimizado con el titular del mismo. Para ello las técnicas más utilizadas son:

- **Cifrado con clave secreta**: el poseedor de la clave puede reidentificar al individuo de una forma sencilla. Con esta técnica de pseudonimización solamente se necesita descifrar el conjunto de datos, ya que este contiene los datos personales.
- **Función hash**: este método utiliza una función que independientemente del tamaño del valor de entrada devuelve un valor de tamaño fijo. Si se conoce el rango de los valores de entrada de dicha función, se le pueden pasar estos valores para obtener el valor real de cualquier individuo. Cabe la posibilidad de crear tablas precalculadas para lograr una reversión masiva de una gran cantidad de valores hash.
- **Función con clave almacenada**: es un tipo de función hash que recibe de forma suplementaria una clave secreta. El responsable del tratamiento de datos puede ejecutar la función con el atributo y la clave secreta.
- **Cifrado determinista**: esta técnica de pseudonimización consiste en generar un número aleatorio a modo de pseudónimo, para cada atributo de la base de datos. Este tipo de pseudonimización reduce el riesgo de que se vinculen los datos personales del conjunto de datos y los datos personales relativos al mismo individuo en otro conjunto de datos en el que se haya utilizado otro pseudónimo.

[1] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu$_{es}$

- **Descomposición en tokens**: este método de pseudonimización es sobretodo utilizado en el sector financiero. Se utiliza para sustituir los números de identificación de las tarjetas por valores no útiles para los atacantes. Suele basarse en la aplicación de mecanismos de cifrado unidireccionales o bien en la asignación mediante una función que no derive directamente de los datos originales.

CAPÍTULO 3
Anonimización de texto

En el presente capítulo, se van a comentar los modelos de lenguaje pre-entrenados empleados en el presente proyecto. A continuación, se van a comentar las técnicas empleadas para la generación de conjuntos de datos anonimizados, así como ejemplos específicos de cada técnica y las estadísticas que se han extraído.

3.1 Modelos de lenguaje pre-entrenados

Como se ha comentado en la sección 2.2, SpaCy es una de las mejores opciones para la tarea de reconocimiento de entidades nombradas, debido a su alta velocidad y la capacidad de personalización de sus modelos. Por lo tanto, tras realizar experimentos con distintos modelos de lenguaje pre-entrenados, se ha escogido, para el desarrollo del presente proyecto, un modelo de lenguaje pre-entrenado basado en SpaCy con *fine-tuning* para el idioma español, llamado *es_core_news_lg*[1].

Figura 3.1: Arquitectura de los modelos SpaCy y su funcionalidad.

[1]https://spacy.io/models/es#es_core_news_lg

3.2 Técnicas de Anonimización

Hay que tener en cuenta que para una posible anonimización de texto primero hay que realizar la tarea de reconocimiento de entidades nombradas. Para realizar esta tarea realizamos tres técnicas diferentes en el presente proyecto (ver 3.2.1, 3.2.2, 3.2.3) [11], una vez reconocidas las entidades, estas de anonimizarán.

3.2.1. Diccionarios

La primera aproximación que se propone en el presente trabajo es la creación de diccionarios, los cuales contienen palabras asociadas a diferentes tags, cada tag esta relacionado con una entidad a anonimizar. Con estos diccionarios creados se emplearán para buscar cada una de las palabras dentro del texto, de tal forma que al encontrarlas se sustituyen por los tags asociados a ellas. La motivación que sigue a esta técnica es la capacidad y facilidad de anonimizar diferentes palabras de forma inequívoca.

Aunque con el método de la utilización de los diccionarios podemos detectar de forma inequívoca cada uno de los elementos de los diccionarios, este método presenta una gran limitación. Esta limitación se puede ver en el primer ejemplo de la tabla 3.4. Al no estar la entidad Comunidad Valenciana añadida en el diccionario, esta entidad no es capaz de anonimizarse, por lo que este método de por si solo no es viable.

Diccionario	Original	NER	Anonimizado
Juan, PER Gestalgar, LOC	**Juan** vive en un pueblo en el interior de la Comunidad Valenciana llamado **Gestalgar**	**PER** vive en un pueblo en el interior de la Comunidad Valenciana llamado **LOC**	___ vive en un pueblo en el interior de la Comunidad Valenciana llamado ___
Roma, LOC César Augusto, PER	Durante los tres siglos anteriores al ascenso de **César Augusto**, **Roma** pasó de ser uno de los tantos Estados de la península itálica a unificar toda la región y expandirse más allá de sus límites.	Durante los tres siglos anteriores al ascenso de **PER**, **LOC** pasó de ser uno de los tantos Estados de la península itálica a unificar toda la región y expandirse más allá de sus límites.	Durante los tres siglos anteriores al ascenso de ___, ___ pasó de ser uno de los tantos Estados de la península itálica a unificar toda la región y expandirse más allá de sus límites.
Unión Europea, ORG Tratado de Maastricht, MISC Europa, LOC	La **Unión Europea** (UE) es una comunidad política de derecho constituida en régimen *sui generis* de organización internacional nacida para propiciar y acoger la integración y gobernanza en común de los Estados y los pueblos de **Europa**. Está compuesta por veintisiete Estados europeos y fue establecida con la entrada en vigor del **Tratado de Maastricht** el 1 de noviembre de 1993.	La **ORG** (UE) es una comunidad política de derecho constituida en régimen *sui generis* de organización internacional nacida para propiciar y acoger la integración y gobernanza en común de los Estados y los pueblos de **LOC**. Está compuesta por veintisiete Estados europeos y fue establecida con la entrada en vigor del **MISC** el 1 de noviembre de 1993.	La ___ (UE) es una comunidad política de derecho constituida en régimen *sui generis* de organización internacional nacida para propiciar y acoger la integración y gobernanza en común de los Estados y los pueblos de ___. Está compuesta por veintisiete Estados europeos y fue establecida con la entrada en vigor del ___ el 1 de noviembre de 1993.

Tabla 3.1: Ejemplos de texto anonimizado a través de dicionarios.

3.2.2. Expresiones regulares

La segunda aproximación que se propone en el presente trabajo es el reconocimiento de entidades mediante el uso de reglas de expresiones regulares. Estas reglas permiten reconocer entidades con un patrón determinado el cual siempre siguen. Por ejemplo, este podría ser el caso de los correos que siempre tiene el mismo formato, un número indeterminado de carácteres antes del @ y conjunto de caracteres seguidos de un punto y un dominio como son *.com, .es, .org*. Como en este caso hay otras entidades que tambien se pueden reconocer mediante las reglas de expresión regular para más tarde anonimizarlas.

Con este método se consiguen mejoras respecto al método de anonimización mediante diccionarios, ya que con una sola regla se recogen muchas entidades de un mismo formato. Esto se puede ver en los ejemplos de la tabla 3.2, donde se utilizan expresiones regulares para el número de teléfono[2], para el IBAN[3] y para el dinero[4]. Aunque con las reglas se consigue una mejora no es suficiente, ya que hay ciertas entidades que no respetan ninguna regla de expresiones regulares, como pueden ser los nombres de personas o los nombres de las organizaciones.

[2] ((?:\\ + |00)[17](? : |\\−)?|(? : \\ + |00)[1 − 9]\\d{0,2}(? : |\\−)?|(? : \\ + |00)1\\ − \\d{3}(? : |\\−)?)?(0\\d|\\([0 − 9]{3}\\)|[1 − 9]{0,3})(? : ((? : |\\−)[0 − 9]{2}){4}|((? : [0 − 9]{2}){4})|((? : |\\−)[0 − 9]{3}(? : |\\−)[0 − 9]{4})|([0 − 9]{7}))|(\\ + \\d{1,3})?\\d{3}\\d{3}\\d{3}|(\\ + \\d{1,3})?\\d{3}\\d{2}\\d{2}\\d{2}|\\d{2}\\d{3}\\d{2}\\d{2}

[3] ([A-Z]{2}[0-9]{2}([A-Z0-9]{4}){2}([A-Z0-9]{3,4}){1,2}([A-Z0-9]{1,7}){0,1}([A-Z0-9]{1,4}){0,1}([A-Z0-9]{1,4}){0,1} | [A-Z]{2}[0-9]{2}[A-Z0-9]{11,28}) | \\d{4} − \\d{3} − \\d{7}

[4] (([0-9]{1,}((, | .)[0-9]{1,2})?((euros | dólares) | ?(€ | $))) | ([0-9]{3}(.[0-9]{3}){0,}(,[0-9]{1,2})?((euros | dólares) | ?(€ | $))))

Expresiones regulares	Original	NER	Anonimizado
Regla teléfono, PHONE	El otro día Juan recibió una llama del número **655452658**, número que él no tenía agregado, también recibió una llamada días antes del **+39 377 682 2688**, un número proveniente de Italia.	El otro día Juan recibió una llama del número **PHONE**, número que él no tenía agregado, también recibió una llamada días antes del **PHONE**, un número proveniente de Italia.	El otro día Juan recibió una llama del número ___, número que él no tenía agregado, también recibió una llamada días antes del ___, un número proveniente de Italia.
Regla cuenta de banco, IBAN	Los códigos IBAN tienen un formato diferente dependiendo del país donde la cuenta de banco está registrada, por ejemplo en España el IBAN empieza por ES y tiene 24 caracteres en total, mientras que en Bélgica tiene el siguiente formato **BE68 5393 6754 7018**, empieza por BE y tiene 16 caracteres.	Los códigos IBAN tienen un formato diferente dependiendo del país donde la cuenta de banco está registrada, por ejemplo en España el IBAN empieza por ES y tiene 24 caracteres en total, mientras que en Bélgica tiene el siguiente formato **IBAN**, empieza por BE y tiene 16 caracteres.	Los códigos IBAN tienen un formato diferente dependiendo del país donde la cuenta de banco está registrada, por ejemplo en España el IBAN empieza por ES y tiene 24 caracteres en total, mientras que en Bélgica tiene el siguiente formato ___, empieza por BE y tiene 16 caracteres.
Regla dinero, MONEY	Hace unos meses Elon Musk quería comprar Twitter por **44.000.000$** peró al final decidió echarse para atrás.	Hace unos meses Elon Musk quería comprar Twitter por **MONEY** peró al final decidió echarse para atrás.	Hace unos meses Elon Musk quería comprar Twitter por ___ peró al final decidió echarse para atrás.

Tabla 3.2: Ejemplos de texto anonimizado a través de expresiones regulares.

3.2.3. Modelos NER

La siguiente técnica de anonimización texto no estructurado que se propone es el uso de modelos de reconocimiento de entidades nombradas, como es el modelo preentrenado de spaCy mencionado en la anterior sección 2.2.1. El modelo NER de spaCy seleccionado, mencionado en 3.1, es un *Transition-Based Parser*. El *Transition-Based Parsing* es una aproximación de la predicción estructurada en el que la tarea de predecir la estructura se asigna a una serie de *state transitions*.

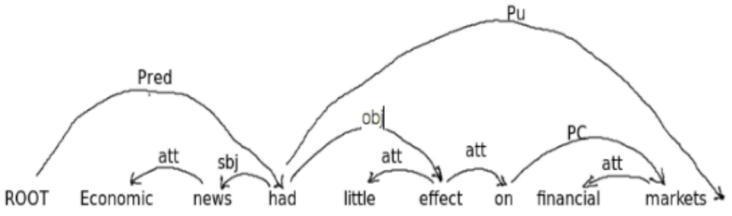

Figura 3.2: Transition-Based Parser.

El modelo *Transition-Based Parser* de la red neuronal consta de dos: la primera subred transforma cada *token* en un vector de representación. La segunda se encarga de construir un vector específico de características para cada par (token, característica), las dos subredes se ejecutan una vez por cada *batch*. De esta forma, al juntar las dos subredes, se crea una red neuronal convolucional, *CNN* [10], con *Bloom Embeddings*.

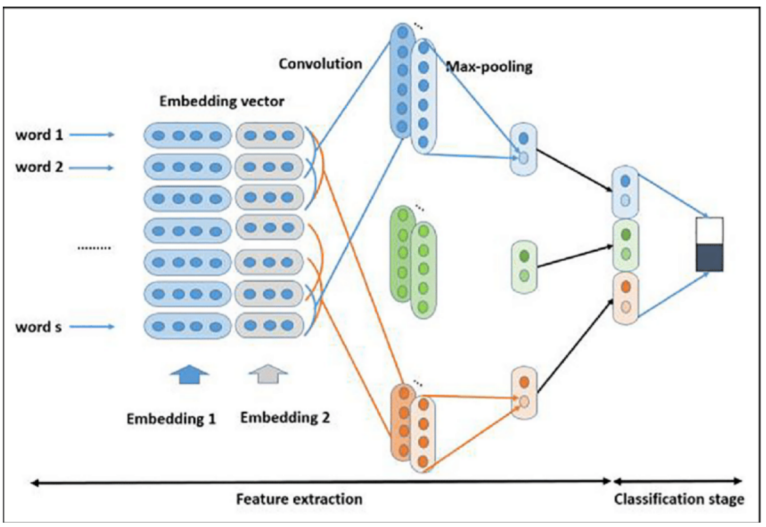

Figura 3.3: Convolutional Neural Network with Embeddings.

Por último, en la tabla 3.3 se muestran algunos ejemplos de la anonimización al aplicar este último método. Vemos una gran mejora ya que no es necesario que introduzcamos cada palabra o entidad como en el caso de los diccionarios 3.2.1 ni tampoco hace falta introducir cada regla 3.2.2. Pese a presentar mejoras no es perfecto, ya que el modelo base de spaCy solo reconoce 4 tipos de entidades, personas, organizaciones, localizaciones y misceláneo (incluye títulos de libros, productos, acontecimientos históricos, etc). La escasez de entidades nos lleva a decidir realizar reentranamientos en el próximo capítulo 4, aparte de definir expresiones regulares para añadir más tipo de entidades.

Original	NER	Anonimizado
Miguel de Cervantes Saavedra (**Alcalá de Henares**, 29 de septiembre de 1547-**Madrid**, 22 de abril de 1616) fue un novelista, poeta, dramaturgo y soldado español, su principal obra fue **El Quijote**.	**PER** (**LOC**, 29 de septiembre de 1547-**LOC**, 22 de abril de 1616) fue un novelista, poeta, dramaturgo y soldado español, su principal obra fue **MISC**.	___ (___, 29 de septiembre de 1547-___, 22 de abril de 1616) fue un novelista, poeta, dramaturgo y soldado español, su principal obra fue ___.
El 1 de abril de 1976 fue fundada **Apple Computer** a través de un contrato firmado por sus tres accionistas: **Steve Wozniak**, **Steve Jobs** y **Ron Wayne**, este último antiguo compañero de trabajo de **Jobs** en la empresa **Atari**, y con apenas 10 % de la nueva empresa.	El 1 de abril de 1976 fue fundada **ORG** a través de un contrato firmado por sus tres accionistas: **PER**, **PER** y **PER**, este último antiguo compañero de trabajo de **PER** en la empresa **ORG**, y con apenas 10 % de la nueva empresa.	El 1 de abril de 1976 fue fundada ___ a través de un contrato firmado por sus tres accionistas: ___, ___ y ___, este último antiguo compañero de trabajo de ___ en la empresa ___, y con apenas 10 % de la nueva empresa.
La Biblia está organizada por dos partes principales; **Antiguo Testamento** (**Tanaj**, libros sagrados canónicos en el judaísmo) y el **Nuevo Testamento** que se enfoca en **Jesucristo** y el cristianismo primitivo.	**MISC** está organizada por dos partes principales; **MISC** (**MISC**, libros sagrados canónicos en el judaísmo) y el **MISC** que se enfoca en **PER** y el cristianismo primitivo.	___ está organizada por dos partes principales; ___ (___, libros sagrados canónicos en el judaísmo) y el ___ que se enfoca en ___ y el cristianismo primitivo.

Tabla 3.3: Ejemplos de texto anonimizado a través del modelo NER de spaCy.

3.3 Tipos de Anonimización

Una vez hemos reconocido y localizado las entidades dentro de texto no estructurado se debe decidir qué hacer con esas entidades, o mejor dicho, de qué forma anonimizar dichas entidades. Dentro de las diferentes opciones que tenemos de anonimizar destacamos cuatro:

- **Eliminación de entidades**: en este tipo de anonimización se opta por eliminar directamente cada una de las entidades del texto. Esta opción no es muy recomendada ya que aunque cumple su cometido, al no ver nada reflejado como una etiqueta o otra cosa que identifique que en ese lugar del texto había algún dato sensible puede llevar a una peor comprensión del documento.
- **Sustitución por etiqueta**: con este tipo de anonimización se consigue reflejar en el texto donde esta toda la información sensible, además de indicar a que dominio pertenece dicha entidad. Esta opción es un paso intermedio hacia la anonimización total del documento, ya que sigue guardando información útil.
- **Sustitución por pseudónimo**: al igual que en el anterior tipo de anonimización se sigue guardando información, ya que al sustituir una entidad por otra del mismo tipo (por ejemplo, Juan por Luís) se consigue ver a que tipo de entidad hace referencia. La virtud de esta opción está en la fácil lectura de estos documentos.
- **Tachado o borrado de entidades**: con esta opción se consigue anonimizar totalmente cada una de las entidades a la vez que mantiene una fácil comprensión del documento. Al contrario que en la eliminación de entidades, ya que con este tipo de anonimización se tacha con un carácter determinado (puede ser "*", "_", "-", o cualquier otro).

Original	Eliminación	Tag	Pseudónimos	Tachado
María y **Juan** están organizando un viaje a **Lisboa**, pero están teniendo problemas para seleccionar la aerolínea ya que no saben si coger **Iberia** o **Air Portugal**.	y están organizando un viaje a , pero están teniendo problemas para seleccionar la aerolínea ya que no saben si coger o .	**PER** y **PER** están organizando un viaje a **LOC**, pero están teniendo problemas para seleccionar la aerolínea ya que no saben si coger **ORG** o **ORG**.	***Rubén*** y ***Natalia*** están organizando un viaje a ***Londres***, pero están teniendo problemas para seleccionar la aerolínea ya que no saben si coger ***Apple*** o ***Renault***.	___ y ___ están organizando un viaje a ___, pero están teniendo problemas para seleccionar la aerolínea ya que no saben si coger ___ o ___.
La ciudad también fue cuna de famosos escritores como **Marco Polo** (1254-1324) (aunque existe un debate sobre el lugar de nacimiento de **Marco Polo** y se propone que haya nacido en la isla de **Korčula** perteneciente a **Croacia**) y su célebre libro **Il Milione**.	La ciudad también fue cuna de famosos escritores como (1254-1324) (aunque existe un debate sobre el lugar de nacimiento de y se propone que haya nacido en la isla de perteneciente a) y su célebre libro .	La ciudad también fue cuna de famosos escritores como **PER** (1254-1324) (aunque existe un debate sobre el lugar de nacimiento de **PER** y se propone que haya nacido en la isla de **LOC** perteneciente a **LOC**) y su célebre libro **MISC**.	La ciudad también fue cuna de famosos escritores como ***Antonio*** (1254-1324) (aunque existe un debate sobre el lugar de nacimiento de ***Antonio*** y se propone que haya nacido en la isla de ***Creta*** perteneciente a ***Valladolid***) y su célebre libro ***El Hombre Lobo***.	La ciudad también fue cuna de famosos escritores como ___ (1254-1324) (aunque existe un debate sobre el lugar de nacimiento de ___ y se propone que haya nacido en la isla de ___ perteneciente a ___) y su célebre libro ___.

Tabla 3.4: Ejemplos de texto anonimizado a través de dicionarios.

3.4 Personalización de los modelos

Por último, vamos a dotar a los modelos de una cierta personalización, para que cualquier persona, empresa u organismo pueda utilizar la anonimización de la forma que más le convenga para tratar sus documentos privados. Para ello, daremos varias opciones al individuo para que seleccione qué quiere anonimizar, o lo que no quiere que se anonimice.

- **Selección de las etiquetas**: con esta opción el usuario podrá decidir los tags (tipos de entidades) que quiere que se anonimicen en el documento. Los tags disponibles son PER (nombre de persona), LOC (localización), MISC (misceláneo), ORG (organización), EMAIL (dirección de correo), IBAN (número de la cuenta del banco) y PHONE (número de teléfono).

  Esta opción puede ser útil cuando se quiera anonimizar un currículum en el que, por ejemplo, el nombre de la persona se quiere que se anonimice, mientras que la organización donde ha realizado los estudios, o los estudios mismos sí se quiere que queden visibles.

- **Tipo de anonimización**: esta opción va relacionada con la anterior sección 3.3, ya que se permite al usuario seleccionar el tipo de anonimización que prefieran. Pese a que en el apartado anteriormente comentado estaba incluida el tipo de anonimización por eliminación, se ha decidido quitarla. Esto se debe a que como ya se ha dicho, puede llevar a una mala comprensión del texto, encontrándote frases sin sentido.

  Esta opción permite que el usuario elija si premiar la legibilidad del documento, seleccionando sustitución por entidades o pseudónimos, o premiar la eliminación total de la información sensible seleccionando el tachado de entidades.

- **Diccionario de anonimización**: tanto está opción como las dos siguientes sirven para crear un modelo híbrido de anonimización entre modelos de reconocimiento de entidades nombradas de spaCy 3.2.3, la creación de diccionarios 3.2.1, y la creación de reglas 3.2.2.

  Con esta opción el usuario podrá elegir palabras que quiera sí o sí que se anonimicen, independientemente del contexto. Esto sirve cuando quieras anonimizar el nombre de un producto para el que el modelo no esta preparado anonimizar. También es útil si ya has encontrado algún error en anteriores ejecuciones que se repite, en este caso se podría añadir la entidad que da error al diccionario, para que de esta forma se anonimice siempre.

- **Diccionario de limpieza**: Con esta opción consigues el efecto contrario a la anterior. Se consigue que las palabras o entidades añadidas a este diccionario pase lo que pase no se anonimicen. Esta opción puede ser útil a la hora de anonimizar documentos de una empresa que tiene el nombre de clientes y empleados. Por ejemplo, en un hospital los nombres de los pacientes deben ser anonimizados, pero el propio hospital puede querer que en sus documentos siga constando el nombre de los doctores que han atendido u operado a dichos pacientes para que quede registrado.

- **Expresiones regulares**: La última opción con la que se cuenta es la creación de expresiones regulares para la anonimización de las mismas. Esta opción es una opción más técnica que no es útil para todo tipo de usuarios, ya que una persona puede no saber hacer una expresión regular para captar el formato de una entidad, o una empresa puede no tener una persona capacitada para ello. De igual forma se podría hacer una petición para crearse dicha expresión regular en caso de que hiciese

falta. Con esta opción se pueden añadir nuevas entidades. de las cuales no dispone el modelo, que se quieran anonimizar, como podría ser una url, o cantidades de dinero.

Aunque todas estas opciones pueden resultar muy útiles, puede llegar a ser muy tedioso volver a seleccionar todos los parámetros cada vez que se quiera anonimizar un documento nuevo. Por ello esta la opción de añadir todos estos parámetros dentro de un perfil que se puede guardar para cada usuario. Gracias a este perfil podrás almacenar todas las funciones que quieras en él, de tal forma que cuando se quiera anonimizar un nuevo documento con las opciones que suele utilizar un usuario, únicamente tendrá que seleccionar el perfil que desea utilizar.

Cabe destacar que todas estas opciones se seleccionan sobre un modelo híbrido que contiene el modelo base de spaCy y tres reglas de expresión regular, las cuales son la del email, la del iban y la del número de teléfono.

CAPÍTULO 4
Marco experimental

En el siguiente capítulo se va a presentar tanto la metodología como las tecnologías empleadas para el desarrollo de los modelos de anonimización, se van a crear nuevos modelos de anonimización a partir de los preentrenados, se evaluarán los resultados y finalmente se comentarán posibles personalizaciones del modelo para que la anonimización quede al gusto de quien quiera utilizarlo.

4.1 Tecnologías empleadas

Python

Python es el principal lenguaje de programación utilizado en el presente trabajo debido a que es un lenguaje dinámico, cuyo objetivo es alcanzar un código legible y fácil de entender. Además, posee una gran variedad de librerías que facilitan el trabajo en el ámbito del procesamiento del lenguaje natural y del aprendizaje automático.

Bash

Por otro lado, se ha empleado, aunque en menor medida, el lenguaje de programación empleado en la terminal de Unix, *Bash*, debido a que a lo largo del desarrollo del trabajo se ha empleado el sistema operativo Ubuntu 20.04. Se ha utilizado para la ejecución de código *python*, para los reentranamientos de los modelos y la creación de los *docker* donde se subían los modelos, a aparte de los comandos básicos de la propia terminal.

Librerías Python

Para el desarrollo del proyecto se han utilizado muchas librerías diferentes, pero las dos principales han sido:

- **SpaCy**: spaCY es una librería de procesamiento del lenguaje como se ha comentado en 2.2.1. Lo que utilizamos en el presento TFG son los modelos pre-entrenados en español, más concretamente, las partes del modelo utilizadas para el reconocimiento de entidades. También utilizamos la librería para realizar reentrenamientos de los modelos anteriormente mencionados.

- **FastAPI**: FastAPI es un framework para construir APIs de forma sencilla y rápida con Python, actualmente es considerado como uno de los frameworks basados en

Python más rápidos. Esta librería se ha utilizado para construir una API donde hacer los *post* y *requests* de forma sencilla.

Git

Git es un sistema de control de versiones distribuido de código abierto. El control de versiones distribuido permite a los desarrolladores descargar un software, realizar cambios y subir la versión que han modificado. En este proyecto se ha utilizado para almacenar los cambios que se iban haciendo para tener la posibilidad de hacer un *rollback* a una anterior versión del proyecto si esto fuera necesario.

Docker

Docker es una plataforma de software que permite crear, probar e implementar aplicaciones rápidamente. Docker empaqueta software en unidades estandarizadas llamadas contenedores que incluyen todo lo necesario para que las aplicaciones se ejecuten, incluidas bibliotecas, herramientas de sistema, código y tiempo de ejecución.

Trello

Trello es un sistema de administración que actúa como interfaz web enfocado en la creación y la administración de tarjetas, listas y tableros que permiten organizar proyectos. En este TFG ha servido de ayuda a la hora de estructurar todas las tareas que había que hacer durante el desarrollo del proyecto.

4.2 Evaluación

Tras el proceso de entrenamiento de los modelos, se tiene que realizar una evaluación de los mismos con el objetivo de valorar la calidad de los textos anonimizados generados. En el presente proyecto, se emplean tres métricas *F-score*, *Recall* y *Precisión*.

El *F-score* mide la precisión que tiene una prueba, se calcula como la media armónica entre la precisión y el recall, estas dos medidas se encargan, utilizando los verdaderos positivos, los falsos positivos y los falsos negativos, de determinar del conjunto de entidades cuántas se ajustan al término buscado, y cuántas de dichas entidades son realmente lo que se quiere extraer. La precisión mide la cantidad de información que es relevante para la consulta, y el recall mide la fracción de información que se ha recuperado con éxito. Usando estas fórmulas se calculan los valores de dichas métricas.

$$F_1 = \frac{P - R}{P + R} \tag{4.1}$$

$$R_{Recall} = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos} \tag{4.2}$$

$$P_{Precision} = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos} \tag{4.3}$$

4.3 Conjunto de datos

En esta sección, se introduce el conjuntos de datos que se ha utilizado a lo largo del desarrollo del presente proyecto. Dichos conjuntos de datos están estructurado a nivel de frase y palabra. Además, se ha realizado un pre-procesado de los datos. Por último, cabe destacar que el conjunto de datos empleado en este proyecto proviene de WikiNER[12], el cual se va a describir a continuación.

4.3.1. Wikiner

Wikiner es un corpus creado a apartir de artículos de la *Wikipedia*, concretamente 7200 artículos en nueve idiomas diferentes (español, francés, italiano, inglés, alemán, portugués, ruso y holandés), aunque en este proyecto solo s e va a utilizar el corpues en español.

El conjunto de datos está estructurado a nivel de frase y en formato *iob*, presentado por L. A. Ramshaw y M. P. Marcus en [14]. Este formato se utiliza para hacer un tag a cada token o palabra, si dicho tag empieza por I significa que es el token se encuentra dentro de un *chunk* a etiquetar, si empieza por B, significa que es el primer token de un *chunk* que va precedido de otro en el anterior token, y si empieza por O significa que no pertenece a ningún *chunk* y por lo tanto no es necesario etiquetar. De esta forma el etiquetado quedaría como en la siguiente tabla 4.1.

Original	**IOB format**
Cuando Diomedes hirió a Eneas Apolo le rescató.	Cuando | CSUBX | O Diomedes | NP | I-PER hirió | VLfin | O a | PREP | O Eneas | NC | I-PER Apolo | NP | B-PER le | PPC | O rescató | VLfin | O . | FS | O
El extenso territorio de la República Argentina está dotado de grandes atractivos turísticos.	El | ART | O extenso | ADJ | O territorio | NC | O de | PREP | O la | ART | I-LOC República | NC | I-LOC Argentina | NP | B-LOC está | VEfin | O dotado | VLadj | O de | PREP | O grandes | ADJ | O atractivos | NC | O turísticos | ADJ | O . | FS | O

Tabla 4.1: Tagging in IOB format.

En este corpus se recogen cuatro etiquetas diferentes a la hora de etiquetar los *chunks*, son PER (Persona), LOC (Localización), ORG (Organización), MISC (Misceláneo). Además de estas etiquetas en la siguiente sección se crean más etiquetas para un reconocimiento de un mayor número de entidades diferentes.

4.4 Preparación de los datos

A la hora de preparar los datos del corpus de *WikiNER*, se ha decidido incorporar nuevos tags al mismo corpus, exactamente se ha decidido añadir los tags PROF (Profesión) y NAT (Nacionalidad).

Para ello, primero hemos creado una lista de entidades correspondientes a dichos tipos de entidad. Para profesión se ha creado una lista de 9026 profesiones incluyendo el masculino y el femenino. Para la nacionalidad se ha creado una lista de 8776 nacionalidades en masculino y femenino también. A parte de estas dos listas se ha decidido crear también una lista adicional para nombres de calles y pueblos españoles en general.

Una vez creadas estas listas, utilizamos las primeras dos para crear las 2 nuevas etiquetas. Esto lo hacemos buscando cada una de las entidades contenidas en estas listas y asociándole el tag de la lista en formato IOB (en este caso siempre se sustituía el tag O por por el tag correspondiente al tipo de entidad empezando por I o B dependiendo de la situación en la que se encontrase el token etiquetado).

La tercera lista se deseó crear a partir de ver que el modelo base de spaCy no detectaba calles de una forma ideal, a parte que había pueblos menos conocidos que tampoco eran detectados por el modelo. De esta forma se decidió sustituir en el 50 % de las ocasiones el chunk o token etiquetado como localización por uno de los elementos de la lista, para así, conseguir un mejor resultado.

Una vez creadas las nuevas etiquetas y reemplazadas algunas entidades antiguas por otras nuevas se tuvo que conseguir pasar del formato al formato de entrenamiento de spaCy. Este formato es una lista que cada elemento es una tupla que hace referencia a cada una de las frases del conjunto de datos. Cada tupla consta de una clave, la cual es la frase y un valor el cual es un diccionario con clave *'entities'* y valor una lista con tuplas, donde cada tupla hace referencia a una entidad de dicha frase. En esta tupla hay 3 valores, la posición de inicio de la entidad, la posición en la que acaba la entidad más uno, y el tag al que se asocia dicha entidad (PER, LOC, ORG, ...). Se puede ver el formato final del conjunto de datos en la tabla 4.2. También se pueden apreciar el la cantidad exacta cada tipo de entidad presente en el corpus en la tabla 4.3

4.5 Reentrenamiento de modelos

Para el reentrenamiento de los modelos hemos utilizado los archivos de configuración de entrenamiento y los comandos de los que dispone spaCy. Los parametros que hemos definido en la configuración del reentrenamiento son los siguientes:

WikiNer IOB format	Final format
Otra | NP | O versión | NC | O la | PPC | O hace | VLfin | O hija | NC | O de | PREP | O **Yaso | NP | I-PER** , | CM | O rey | NC | O de | PREP | O la | ART | O ciudad | NC | O . | FS | O	('Otra versión la hace hija de **Yaso** , rey de la ciudad . ', {'entities': **[(29, 34, 'PER')]**})
Según | NP | O la | ART | O mitología | NC | O **griega | ADJ | I-PROF** , | CM | O **Rea | NP | I-PER** ocultó | VLfin | O a | PREP | O **Zeus | NP | I-PER** en | PREP | O el | ART | O **monte | NC | I-LOC Ida | NC | I-LOC** , | CM | O situado | VLadj | O en | PREP | O el | ART | O centro | NC | O de | PREP | O la | ART | O isla | NC | O . | FS | O	('Según la mitología **griega** , **Rea** ocultó a **Zeus** en el **monte Ida** , situado en el centro de la isla . ', {'entities': **[(19, 29, 'NAT'), (32, 36, 'PER'), (46, 52, 'PER'), (59, 80, 'LOC')]**})

Tabla 4.2: Formato final.

	Entities	
	WikiNER Original	**WikiNER with additions**
PER	71094	71094
LOC	119854	119854
ORG	21435	21435
MISC	32980	32980
PROF	9026	9026
NAT	8776	8776

Tabla 4.3: Cantidad de entidades en el corpus.

- **Modelo base**: a la hora de seleccionar un modelo a la hora de reentrenar hemos elegido escoger como modelo base el modelo *es_core_news_lg*. Pese a ser muy pesado en comparación a las alternativas, *es_core_news_sm* y *es_core_news_md*, tiene una mejor predicción según las evaluaciones que aparecen en la web de spaCy.

- **Convolutional Neural Network**: como ya se ha explicado anteriormente en la sección 3.2.3, los modelos NER de spaCy tienen están entrenado con una red convolucional. Por esto, se ha seleccionado este tipo de red neuronal para el reentrenamiento de los modelos.

  A la hora de definir todas las características de la red, se ha decidido establecer los siguientes valores:

  - Una **profundidad** de 8 capas.
  - Una **anchura** de 256 neuronas.

- Una **función *maxout*** en la capa de salida de 3 unidades.

- **Optimizador**: Para optimizador se ha elegido el Descenso por Gradiente Estocástico (o en inglés *Stochastic Gradient Descent*, SGD [9]) con disminución de pesos. Se ha establecido el *learning rate* en 0.001, mientras que el factor de regularización *L2*, se ha inicializado en 10^{-6}.

  La actualización de los pesos se lleva acabo mediante las siguientes formulas 4.4 y 4.5, donde la W es la matriz de pesos, ρ es el *learning rate*, G el gradiente, $\frac{dL}{dW}$ es el factor de regularización *L2* [5] y λ el *weight decay*.

$$W^{(i+1)} = W^{(i)} - \rho G^{(i)} \tag{4.4}$$

$$G^{(i)} = \frac{dL}{dW^{(i)}} - \lambda W^{(i)} \tag{4.5}$$

- **Otros parámetros**:

  - El modelo se entrena en ***batches*** de 1000 datos.

  - Se establece un **dropout** [19] de 0.2 para evitar un posible sobreentrenamiento.

  - Se establece el **máximo de pasos** en 30000.

  - El modelo se evalúa cada 200 pasos y guarda el mejor hasta ese momento.

4.6 Experimentación

Para el presente proyecto vamos a realizar 3 experimentos diferentes evaluando el mejor tipo de datos para entrenar modelos dependiendo de si el texto está en mayúscula todo, solo en mayúscula la *capital letter* o es un texto mixto.

Realizamos estos experimentos ya que hay muchos documentos que pueden contener el texto en mayúscula como en un formulario y documentos en le que dentro del texto aparezcan solo algunas palabras en mayúscula (que suelen hacer referencia a alguna entidad). Estos dos casos, con un único modelo entrenado con texto en el que las únicas mayúsculas sean las *capital letters*, serían difíciles de anonimizarse debidamente.

También hay que añadir que cada uno de los modelos, pese a entrenarse con un tipo de texto diferente, provienen del propio corpus WikiNER y se dividen el 90 % en conjunto de entrenamiento y 10 % en conjunto de evaluación.

Para los entrenamientos con texto con *capital letters* no ha hecho falta ninguna modificación respecto al conjunto de datos preparados de WikiNER en la sección 4.4, ya que tiene el formato correcto.

Para los entrenamientos de los modelos con texto solo en mayúscula se ha tenido que modificar los datos. La modificación de los datos consta de realizar un *uppercase* para volver todo el texto a mayúscula.

Para el último experimento, que es con texto mixto (que puede presentar alguna palabras completamente en mayúscula dentro del texto), también hemos tenido que modificar los datos. Para ello, se ha decidido que el 30 % de los datos, tanto de entrenamiento como de evaluación, se les realice un *uppercase* como en el caso anterior, mientras que al otro 70 % no se le ha realizado ningún cambio. Se ha decidido esta división debido a que la presencia de algunas palabras o frases completamente en mayúscula dentro de un texto es menor que la de las palabras o frases con *capital letters*.

CAPÍTULO 5
Resultados

En este capítulo se van a exponer los resultados obtenidos de los distintos modelos de anonimización que se han entrenado. Dichos modelos, se van a evaluar mediante 3 métricas escogidas por su gran capacidad de evaluación.

Antes de pasar a los resultados de los modelos vamos a comprobar como de bueno es el modelo base de spaCy *es_core_news_lg* en la siguiente tabla. Cabe destacar que este modelo no se ha entrenado con las dos últimas etiquetas, por lo que no será capaz de predecirlas.

	Precision	**Recall**	**F1-score**
PER	56.77	65.01	60.61
LOC	35.57	46.83	40.43
ORG	30.15	12.59	17.76
MISC	52.11	80.71	63.33
PROF	0.00	0.00	0.00
NAT	0.00	0.00	0.00

Tabla 5.1: Evaluación del modelo base.

5.1 Entrenamiento con capital letters

En primer lugar, se van a presentar los resultados obtenidos del análisis del modelo entrenado con esxto en *capital letter*. Así, se pueden observar en la tabla 5.2 los resultados obtenidos al aplicar las distintas métricas de evaluación de cada uno de los tipos de entidades del modelo de anonimización.

	Precision	**Recall**	**F1-score**
PER	85.44	91.90	88.55
LOC	94.66	93.14	93.89
ORG	91.69	91.01	91.35
MISC	85.95	80.71	83.24
PROF	77.24	44.61	56.55
NAT	88.05	91.56	89.77

Tabla 5.2: Evaluación del modelo entrenado con datos en capital letter.

A la vista de los resultados vistos en la tabla 5.2, observamos una gran mejora en todas las etiquetas ya definidas en el modelo base. Respecto a las nuevas etiquetas, se puede destacar que la etiqueta *NAT* consigue un muy buen score. En cambio, vemos que la etiqueta *PROF* recibe un score de 56.55 %, aunque cabe destacar que su precisión es bastante buena con un 77.24 %.

5.2 Entrenamiento con mayúsculas

En segundo lugar, se presentan los resultados obtenidos en la evaluación del modelo entrenado con texto completamente en mayúscula. Se pueden observar en la tabla 5.3 los resultados obtenidos al aplicar las distintas métricas de evaluación de cada uno de los tipos de entidades del modelo de anonimización.

	Precision	**Recall**	**F1-score**
PER	52.69	50.62	51.63
LOC	80.15	65.30	71.96
ORG	70.52	57.90	63.59
MISC	70.45	30.64	42.70
PROF	5.83	3.22	4.52
NAT	10.20	15.38	12.79

Tabla 5.3: Evaluación del modelo entrenado con datos en mayúscula.

A la vista de los resultados vistos en la tabla 5.3, observamos que ha mejorado la predicción respecto al modelo base dependiendo de las etiquetas. Por ejemplo, con las etiquetas *LOC* y *ORG* se ven bastante mejoradas respecto al modelo base. En cambio, las etiquetas *PER* y *MISC* su predicción se reduce, en especial baja el score de la etiqueta *MISC*, la cual se reduce en 40 puntos. Respecto a las nuevas etiquetas, se observa que pese a tener un score no nulo, no serían viables añadirlas al modelo, en este caso, por su muy bajo score.

5.3 Entrenamiento con texto mixto

Por último, se presentan los resultados obtenidos en el análisis del modelo entrenado con texto mixto, es decir con datos tanto en mayúscula, como con datos con *capital letters*. Así, se pueden observar en la tabla 5.4 los resultados obtenidos al aplicar las diferentes métricas de evaluación para cada uno de los tipos de entidades del modelo de anonimización.

	Precision	**Recall**	**F1-score**
PER	62.68	61.35	62.01
LOC	80.15	72.34	76.26
ORG	75.60	63.79	69.71
MISC	73.81	60.47	67.13
PROF	15.12	12.97	14.02
NAT	30.67	28.32	29.45

Tabla 5.4: Evaluación del modelo entrenado con datos en texto mixto.

A la vista de los resultados vistos en la tabla 5.4, observamos que ha mejorado la predicción en todas las etiquetas respecto al modelo base. Aunque las mayores mejoras se encuentran en las etiquetas *LOC* y *ORG*, mejorando un 36 % y un 52 % respectivamente, la novedad respecto al modelo anterior es la mejora de las etiquetas *PER* y *MISC*, aunque de forma ligera. Respecto a las nuevas etiquetas, se puede observar que mejora el modelo de la anterior sección, pero el score de la etiqueta *PROF* es demasiado bajo, y en el caso de la etiqueta *NAT* pese a haber mejorado, sigue siendo un 30 % poco útil.

CAPÍTULO 6
Conclusiones

En este último apartado del proyecto, se van a exponer la evaluación de los objetivos que se propusieron al principio del escrito. Además, se va a definir las distintas líneas que se podrían seguir para el desarrollo futuro del presente trabajo.

6.1 Conclusiones

En el presente trabajo, se ha creado una herramienta de anonimización que utiliza diferentes técnicas para tener una anonimización más completa. Por una parte, se han generado 3 modelos diferentes para la anonimización automática empleando un modelo base pre-entrenado con el objetivo de mejorar este modelo. También se ha implementado la anonimización por diccionarios y la anonimización por reglas, aunque estas dos se quedan como opción para el usuario.

Por la parte de los modelos, hemos comprobado que para realizar una correcta anonimización de texto no estructurado, la mejor forma de realizarla es con texto en *capitall letters*. Las opción de los modelos todo en mayúsculas se descarta debido a que incluso el modelo empeora en muchas circunstancias. En cambio, la opción de texto mixto sí que presenta algunas mejoras y podría ser útil en determinados casos. Respecto a los nuevos tags añadidos comprobamos que el único modelo funcional sería el texto con *capital letters*, en el resto de modelos no son una opción viable de momento. Esto se puede deber al desbalanceo de etiquetas, ya que estas dos etiquetas solo tienen 9000 entidades aproximadamente en el corpus y la siguiente que menos tiene es ORG con 21000.

Por último, hemos dotado a la herramienta deuna alta capacidad de personalización, no solo dejándoles poder crear expresiones regulares y diccionarios para anonimizar, sinó otras técnicas como elegir las etiquetas exactas que se quieren anonimizar o el tipo de anonimización a realizar, reflejadas en el 3.4.

6.2 Trabajo futuro

Una primera línea futura, podría ser el reconocimiento y anonimización de imágenes. En el presente proyecto nos centramos únicamente en el texto plano, que puede venir por documentos *.docx*, *.pdf* o *.odt*, entre otros. Esto nos lleva a que surjan problemas respecto a anonimizar una imagen, ya que esta puede llevar texto incluido en ella con información sensible, como podría ser el logo de una empresa. Eso nos llevaría a implementar un *OCR*.

Otra de las vertientes a trabajar, que a la vez está relacionada con la anterior, es la anonimización de las caras de las personas o matrículas en imágenes. Para realizar esto habría que realizar un clasificador de imágenes.

También hay que decir que el presente trabajo se ha centrado sólo en el idioma español, así que se podría realizar este mismo proceso para diferentes idiomas. Gracias a ello, podríamos conseguir llegar a una mejora de los modelos y a una personalización mayor para el usuario en cualquiera de los idioma.

Por último, se pueden implementar nuevas expresiones regulares para anonimizarlas y que así, el usuario las pueda seleccionar sin necesidad de crearlas. Algunas de las nuevas expresiones que se podrían incluir serían una matrícula de coche, una dirección url o documentos identificativos como el DNI o el pasaporte.

Bibliografía

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Masayuki Asahara and Yuji Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 8–15, 2003.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. .ºReilly Media, Inc.", 2009.

[4] Debajit Datta, Preetha Evangeline David, Dhruv Mittal, and Anukriti Jain. Neural machine translation using recurrent neural network. *International Journal of Engineering and Advanced Technology*, 9(4):1395–1400, 2020.

[5] José Ignacio Baciero Fernández. Elaboración de un modelo de reconocimiento de entidades nominales (ner) para su uso en aplicaciones de procesamiento del lenguaje natural (nlp). Junio 2020.

[6] Yolanda Cano Galán. La seudonimización y la anonimización de datos personales en las sentencias del orden jurisdiccional social. *Documentación Laboral*, (119):31–56, 2020.

[7] Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, 2016.

[8] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[9] Kenji Kawaguchi and Haihao Lu. Ordered sgd: A new stochastic optimization framework for empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 669–679. PMLR, 2020.

[10] Miguel Ángel Medina Ramírez. Inteligencia artificial aplicada a la ley de protección de datos. B.S. thesis, 2020.

[11] Salman Naseer, Muhammad Ghafoor, Sohaib, Sohaib Khalid Alvi, Anam Kiran, Shafique Ur Rehman, Ghulam Murtaza, Jehlum Campus, and Pakistan Jehlum. Named entity recognition (ner) in nlp techniques, tools accuracy and performance. 01 2022.

[12] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. Learning multilingual named entity recognition from Wikipedia. 10 2017.

[13] Silvia Olmos Lara et al. Aplicación de la técnica de k-anonimización sobre bases de datos relacionales. 2021.

[14] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999.

[15] Lisa F Rau. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, pages 29–30. IEEE Computer Society, 1991.

[16] Satoshi Sekine. Nyu: Description of the japanese ne system used for met-2. In *Proc. of the Seventh Message Understanding Conference (MUC-7*. Citeseer, 1998.

[17] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*, pages 1977–1980. Lisbon, Portugal, 2004.

[18] José Manuel Simón Ramos et al. Implementación de una herramienta basada en pln para la detección y anonimización de datos personales en documentos. 2021.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

APÉNDICE A
OBJETIVOS DE DESARROLLO SOSTENIBLE

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

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

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

El presente trabajo fin de grado está relacionado con varios de los objetivos de desarrollo sostenible (ODS). Desde el inicio de los tiempos, la comunicación ha supuesto un intercambio de conocimiento y recursos entre distintas sociedades, con el objetivo de socializar y realizar negociaciones. Estos conocimientos y recursos han ido evolucionando en el tiempo hasta convertirse a día de hoy en datos. Estos datos están en todos lados, en forma de texto, imágenes u otras. Pero todos estos datos guardan información sensible que puede ser peligrosa en manos mal intencionadas. Por ello la anonimización automática se encarga de proteger esa información, brindando seguridad cada individuo u organización.

Por otro lado, la anonimización de texto, permite la protección de datos que, dependiendo de las preferencias del usuario, partiendo de un único texto se pueden realizar diferentes anonimizaciones con la personalización del modelo. Así, es posible crear anonimizaciones con diccionarios consiguiendo que las anonimzaciones sean más inclusivas, diversificadas y que tengan en cuenta a una inmensa cantidad de personas, culturas, países, géneros u orientaciones sexuales.

Por último, el presente trabajo supone un gran impacto en la industria, en concreto en del lenguaje, e innovación, debido a que consigue mejorar la anonimización en español con el empleo de técnicas de reconocimiento de entidades, lo que posibilita una mayor privacidad y seguridad. Además, el crecimiento económico también se ve influenciado, ya que al conseguir mejores anonimizaciones, se consigue ahorrar en posibles post-procesos de corrección humana.