

Grado en ingeniería informática
Curso 2022-2023

Trabajo Fin de Grado

“Análisis de literatura científica en el
marco de los Objetivos de Desarrollo
Sostenible mediante un modelo
clasificador de textos basado en Deep
Learning y técnicas de web scraping”

Ignacio Aguilera Cárdenas

Tutora
Ascensión López Vargas
Leganés, Septiembre 2023



Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento - No Comercial - Sin Obra Derivada**

RESUMEN

Los Objetivos de Desarrollo Sostenible son, hoy en día, un pilar fundamental del desarrollo humano ya que estos definen las diversas áreas de interés y los esfuerzos necesarios para llevar a cabo un desarrollo humano que pueda prolongarse en el tiempo y que sea justo para todos. Entender de una manera profunda estos objetivos y su papel en la sociedad es fundamental si se quieren cumplir las metas de la agenda 2030.

Este trabajo expone diversos modelos de aprendizaje profundo desarrollados y sus arquitecturas con el objetivo de clasificar textos acorde a los objetivos con los que están relacionados, explicando las diferentes arquitecturas y datos y exponiendo los resultados y validaciones de los mismos.

De manera adicional se explora la posibilidad de adaptar modelos de clasificación diseñados para asignar una única clase para resolver problemas de clasificación multi-etiqueta.

Finalmente se aplica el mejor de los modelos desarrollados en un ambiente de estudio real, clasificando multitud de textos científicos sobre sostenibilidad y exponiendo de diferentes maneras los resultados obtenidos.

Palabras clave: Aprendizaje Automático; Aprendizaje Profundo; Inteligencia Artificial; Procesamiento del Lenguaje Natural; Aprendizaje por Transferencia; Web Scraping; Objetivos de Desarrollo Sostenible; Sostenibilidad; Agenda 2030

ABSTRACT

The Sustainable Development Goals are nowadays a key aspect of human development, defining the various areas of interest along with the necessary efforts to undertake in order to achieve and secure human development for the foreseeable future. Grasping a deep understanding of these goals and their role in society is a crucial part of accomplishing the 2030 agenda.

This project exposes several deep learning models and their architectures, created with the objective of classifying texts according to the goals they are related to. Meanwhile explaining the how and why of the different architectures, datasets and exposisng and explaining the results and validation of those models.

Additionally the possibility of adapting models designed to classify one label at a time to solve multi-label problems is explored.

Finally the best model is used in a real wold study application, classifying several scientific publications related to sustainability and exposing, in a few different ways, the results obtained from said classification.

Key words: Machine Learning; Deep Learning; Artificial Intelligence; Natural Language Processing; Transfer Learning; Web Scraping; Sustainable Development Goals; Sustainability; Agenda 2030

DEDICATORIA

Este proyecto significa el fin de una de las etapas más importantes en la vida de cualquier persona que tome el camino de la formación universitaria. Con el se da pie a la vida realmente adulta.

Es aquí donde quiero agradecer a todas aquellas personas que me han acompañado durante todos estos años en lo que, habitualmente, es la mejor etapa de la vida de una persona. A aquellas que aún siguen acompañándome y a aquellas que la vida les ha llevado por otro camino, todas ellas forman parte fundamental de mis logros y de quien soy.

Gracias a Juanma y Jaime por todos esos meses de biblioteca y tertulia, seguimos sin saber como conseguimos aprobar absolutamente todo.

Gracias a Paula por estar ahí para mi durante todos estos años, por apoyarme como la que mas y por quererme tantísimo. Gracias también al resto de Iguanas por ser el mejor grupo de amigos que uno podría pedir.

Gracias a Fabi por aparecer en mi vida durante esta etapa, darme tanto amor, aguantarme y ayudarme a aprobar cálculo diferencial.

Gracias al resto de compañeros que ahora puedo afirmar que son mis amigos de verdad. Gracias por hacer tan divertidas las clases y por escaparnos a la playa o al pueblo cada vez que teníamos la oportunidad.

También darle las gracias a Asu por acceder a tutorizarme y guiarme en el desarrollo de este proyecto, espero no haberte decepcionado.

Finalmente gracias a mi familia, a Lucia, a Jose, a Isa y a Tomás, por criarme y por apoyarme incondicionalmente y poner interés en lo que estaba haciendo, aunque a veces resulte complicado de entender.

ÍNDICE GENERAL

1. INTRODUCCIÓN	1
1.1. Introducción al tema	1
1.2. Problema	1
1.3. Motivación	2
1.4. Objetivo	2
1.5. Marco regulador	3
1.6. Estructura del documento	4
2. ESTADO DE LA CUESTIÓN	6
2.1. Objetivos de Desarrollo Sostenible	6
2.2. Inteligencia artificial	8
2.2.1. Historia de la IA	9
2.2.2. Aprendizaje por transferencia	13
2.3. Procesamiento del Lenguaje Natural	15
2.4. Clasificación multi etiqueta	18
2.4.1. Función de activación	19
2.5. Web Scraping	20
2.6. Tecnologías usadas	20
2.7. Trabajos similares	21
3. SOLUCIÓN PROPUESTA/MÉTODO	24
3.1. Arquitectura del sistema	24
3.1.1. Recogida de datos	25
3.1.2. Gestión de datos	28
3.1.3. Gestión de modelos	30
3.2. Desarrollo realizado	33
4. VALIDACIÓN Y PRUEBAS	39
4.1. Conjunto de pruebas	39
4.1.1. Métricas redes recurrentes	42
4.1.2. Métricas modelos BERT	43

4.1.3. Métricas finales	53
4.1.4. Conclusiones parciales	55
4.2. Textos seleccionados	55
4.2.1. Textos Sostenibilidad y sanidad	56
4.2.2. Textos Sostenibilidad e industria	57
4.2.3. Textos Sostenibilidad e igualdad	59
4.2.4. Textos Sostenibilidad y energía	60
4.2.5. Conclusiones parciales	61
4.3. Literatura científica	62
4.3.1. Modelos de redes recurrentes	62
4.3.2. Modelos de BERT	63
4.3.3. Análisis de los datos	67
4.3.4. Conclusiones parciales	85
4.4. Conclusiones generales	85
5. RESULTADOS	87
5.1. Análisis de resultados generales	88
5.2. Evolución de publicaciones	89
5.3. Proporciones generales	95
6. GESTIÓN DEL PROYECTO	96
6.1. Planificación	96
6.1.1. Presupuesto:	98
6.2. Impacto socio-económico	99
7. CONCLUSIONES	102
7.1. Conclusiones generales	102
7.2. Dificultades y limitaciones	102
7.3. Líneas de trabajo futuro	103
ACRÓNIMOS	104
GLOSARIO	106
BIBLIOGRAFÍA	111

ÍNDICE DE FIGURAS

2.1	Modelo célula McCulloch Pitts	10
2.2	Representación perceptrón multicapa	11
2.3	Representación red convolucional	11
2.4	Representación red recurrente	12
2.5	CBOW vs SG	17
2.6	Clasificación números	18
3.1	Diagrama UML de la arquitectura de recogida de datos	25
3.2	Diagrama UML de la gestión de la base de de datos	28
3.3	Diagrama UML de la gestión de modelos	30
4.1	Métricas modelo 1	42
4.2	Métricas modelo 2	43
4.3	Métricas modelo 11	44
4.4	Métricas modelo 13	44
4.5	Métricas modelo 14	45
4.6	Métricas modelo 16	46
4.7	Métricas modelo 17	46
4.8	Métricas modelo 18	47
4.9	Métricas modelo 19	48
4.10	Métricas modelo 20	48
4.11	Métricas modelo 21	49
4.12	Métricas modelo 21	50
4.13	Métricas modelo 23	51
4.14	Métricas modelo 24	51
4.15	Métricas modelo 25	52
4.16	Métricas modelo 26	52
4.17	Exactitud de todos los modelos	53
4.18	Valor-F1 de todos los modelos	54

4.19	Mapa de relaciones del objetivo 8	69
4.20	Mapa de relaciones del objetivo 9	71
4.21	Mapa de relaciones del objetivo 10	73
4.22	Mapa de relaciones del objetivo 11	75
4.23	Mapa de relaciones del objetivo 12	77
4.24	Mapa de relaciones del objetivo 13	79
4.25	Mapa de relaciones del objetivo 14	80
4.26	Mapa de relaciones del objetivo 15	82
4.27	Mapa de relaciones del objetivo 16	83
4.28	Mapa de relaciones del objetivo 17	84
5.1	Resultados generales	89
5.2	Evolución ODS 12	90
5.3	Evolución ODS 3	90
5.4	Evolución ODS 9	91
5.5	Evolución ODS 2	92
5.6	Evolución ODS 15	92
5.7	Evolución del resto de objetivos	94
5.8	Porcentaje de artículos por ODSa	95
6.1	Diagrama de Gantt	98

ÍNDICE DE TABLAS

4.1	Número de datos por objetivo	41
4.2	Métricas estadísticas de la base de datos final	41
4.3	Porcentajes texto 1 - sanidad	56
4.4	Porcentajes texto 2 - sanidad	57
4.5	Porcentajes texto 3 - sanidad	57
4.6	Porcentajes texto 1 - industria	58
4.7	Porcentajes texto 2 - industria	58
4.8	Porcentajes texto 1 - igualdad	59
4.9	Porcentajes texto 2 - igualdad	59
4.10	Porcentajes texto 1 - energía	60
4.11	Porcentajes texto 2 - energía	61
4.12	Porcentajes texto 3 - energía	61
4.13	Media de clasificaciones por modelo recurrente	62
4.14	Clasificaciones de cada modelo recurrente por objetivo	63
4.15	Clasificaciones modelos softmax	64
4.16	Clasificaciones modelos sigmoid - datos iniciales	65
4.17	Clasificaciones modelos sigmoid - datos aumentados	66
4.18	Clasificaciones modelos sigmoid - datos extraídos de internet/finales	67
4.19	Clasificaciones asbtracts ODS8 - modelo 23	70
4.20	Clasificaciones asbtracts ODS9 - modelo 23	72
4.21	Clasificaciones asbtracts ODS10 - modelo 23	74
4.22	Clasificaciones asbtracts ODS11 - modelo 23	76
4.23	Clasificaciones asbtracts ODS12 - modelo 23	78
4.24	Clasificaciones asbtracts ODS14 - modelo 23	81
4.25	Clasificaciones asbtracts ODS17 - modelo 23	85
5.1	Datos resultantes del estudio	88
6.1	Tabla costes de personal.	98

6.2	Tabla costes de software y hardware.	99
6.3	Tabla costes indirectos.	99
6.4	Tabla costes totales.	99

1. INTRODUCCIÓN

Esta sección tiene como objetivo definir las bases contextuales y motivadoras detrás del proyecto, de manera adicional se presenta el marco regulador que lo enmarca y una enumeración de las diferentes secciones que componen el resto del documento.

Destacar que el contexto sobre el que se basa el proyecto son las publicaciones científicas, es por esto por lo que se incluirá y hará referencia a multitud de textos en inglés. Esto es debido a que esta es la lengua predominante en las publicaciones científicas de alto impacto.

1.1. Introducción al tema

Hoy en día la humanidad se enfrenta al que posiblemente sea el problema más difícil de su larga historia, frenar de forma contundente y efectiva los efectos del cambio climático. Este, ocasionado de forma directa por la actividad humana, descontrolada y sin ningún tipo de regulación, ignorando las múltiples señales y consecuencias que vienen aconteciendo durante años, se presenta como una fuerza imparable que está empezando a mostrar la cruda realidad que aguarda.

Para conseguir efectos paliativos significativos sobre las consecuencias del cambio climático es esencial una colaboración global sin precedentes, concienciando a todas las naciones para centrar el esfuerzo de manera colectiva y organizada para conseguirlo. Es por esto, y con esta meta común en el punto de mira, por lo que las Naciones Unidas como institución, formalizaron los Objetivos de Desarrollo Sostenible (ODS), poniendo así sobre el papel una serie de directivas a seguir para asegurar un futuro digno para las próximas generaciones.

El que seguramente sea el esfuerzo fundamental a la hora alcanzar cualquier meta que la sociedad como conjunto se proponga es el constante desarrollo de un entorno investigador y científico activo y eficaz, que se encargue de guiar el progreso en materia de conocimiento por el camino correcto, desarrollando tecnologías relevantes a la vez que sostenibles.

1.2. Problema

Los ODSs, una vez se presta la atención necesaria, están presentes en casi todos los aspectos del día a día. Las corporaciones e instituciones más grandes los llevan por bandera. Afirmando y publicitando multitud de esfuerzos y políticas desarrolladas con estos en mente. Estos esfuerzos deben ser monitorizados para confirmar o desmentir lo que estas pregonan.

El principal problema a la hora de monitorizar el progreso hacia el cumplimiento de los objetivos es la dificultad de analizar la cantidad, sin precedentes, de información generada de manera continua. Esta, teniendo una naturaleza heterogénea y siendo el lenguaje natural, punto central de este trabajo, por naturaleza, variable y costoso de analizar, crea una necesidad latente de procesar, catalogar y entender de manera automática y fiable textos de este carácter.

1.3. Motivación

El monitoreo constante de los esfuerzos que se llevan, o no, a cabo por cumplir dichos objetivos es una tarea esencial y, que como todo buen monitoreo, debe de realizarse de manera independiente y autónoma. Esta tarea es del tipo perfecto para ser automatizada por una serie de técnicas muy variadas. Una de estas técnicas, cada vez más popular, es el aprendizaje automático, y más concretamente el aprendizaje profundo.

El modelo de aprendizaje profundo desarrollado en este trabajo tiene, como motivación, proporcionar la capacidad de analizar textos de manera fiable y automática, habilitando la posibilidad de realizar análisis cuantitativos, basándose en un conjunto de datos cualitativos y, en cierta medida, subjetivos, ya que la presencia de un tema en concreto en un texto es una decisión puramente subjetiva, naciendo esta del conocimiento almacenado por cada individuo y su experiencia, vivencia e intereses pueden afectar de manera significativa a esta decisión.

Eliminar este factor humano que proporciona puntos de conflicto es, con la realización de un estudio cuantitativo y objetivo en mente, un requisito fundamental que cumplir si se quiere legitimar los resultados.

1.4. Objetivo

El objetivo principal que engloba este proyecto es el poder clasificar textos de acuerdo a los ODS relacionados con el mismo. Esta tarea, para poder llevarse a cabo de una manera correcta requiere de una serie de metas más específicas, la primera de las cuales es la extracción de datos de internet relacionados con los ODS y etiquetados de acuerdo a estos objetivos presentes en el texto. En segundo lugar se encuentra el diseño y entrenamiento de un modelo de aprendizaje profundo que sea capaz de extraer la información almacenada en los datos recolectados e identificar los diferentes patrones para que posteriormente, dado un texto nuevo, sea capaz de identificar los diferentes objetivos relacionados con el mismo. Como última tarea está la validación de dicho modelo, esta tiene como finalidad llevar a cabo una serie de pruebas controladas sobre el modelo para analizar si su rendimiento es el adecuado y puede ser usado de manera fiable en tareas de análisis de datos.

Como objetivo adicional está el llevar a cabo un estudio cuantitativo sobre un corpus

textual extraída de bases de datos académicas en base a las clasificaciones generadas por el modelo. De esta manera se estará usando el modelo desarrollado en un entorno real a la vez que se lleva a cabo un estudio que puede arrojar luz sobre el panorama científico actual y como se están distribuyendo los esfuerzos.

Adicionalmente el trabajo realizado para el desarrollo de este modelo, estando este publicado en *github* [1]¹, puede ser usado como referencia para desarrollar modelos clasificadores de textos en otros ámbitos. Experimentando con datos y modelos de diferentes tipos.

1.5. Marco regulador

El trabajo desarrollado consta de un modelo clasificador de textos cuya base de conocimiento ha sido extraída de internet y tiene como objetivo clasificar una serie de textos provenientes de bases de datos académicas como scopus. Es esta extracción de datos el principal aspecto del proyecto que debe adecuarse a las regulaciones. En el caso de España, lugar donde se ha realizado dicho proyecto, aplica de manera directa toda aquella legislación y normativa recogida en el marco jurídico español. Aún así este se amolda a la legislación europea, que, por medio de directivas y reglamentos guían a las instituciones jurídicas de los estados miembros por un camino común hacia la creación de un marco regulador en cierta medida homogéneo en el entorno europeo.

Como primer punto a comentar, el uso de técnicas de web scraping y el posterior tratamiento de los datos extraídos, con el objetivo de entrenar modelos de Aprendizaje Automático, particularmente en el ámbito de clasificación de textos relacionados con los Objetivos de Desarrollo Sostenible, debe ajustarse a las regulaciones existentes relacionadas con la propiedad intelectual, protección de datos y derechos de autor. En España, esto se rige bajo varias leyes y regulaciones, detalladas a continuación:

- 1. Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD):** Esta ley, aunque centrada principalmente en la protección de datos personales, es relevante en cuanto a la extracción de datos de internet ya que puede ser extraída información personal. Aunque la extracción de información de esta índole se haga de manera no intencionada se estaría incumpliendo la ley, por lo que el correcto filtrado y trato de información personal es requerido.[2]
- 2. Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia.:** Este real decreto enmarca toda la legislación española relacionada con la propiedad intelectual, definiendo los derechos de autor, la protección que estos derechos le brindan a las obras

¹La disponibilidad de una versión actualizada no está garantizada debido a la dificultad de adjuntar modelos de alto tamaño

y las excepciones y limitaciones de dichos derechos. Cualquier proyecto que trate con datos tiene que regirse a la normativa expuesta en este decreto.[3]

3. **Legislación de la UE:** En harmonía con la legislación europea, el marco jurídico español se ha amoldado de tal forma que incorpora aspectos provenientes de instituciones europeas en su legislación propia. Esta legislación europea tiene dos formatos, los reglamentos y las directivas. Los reglamentos son directamente aplicables a todos los estados miembros de manera directa, por otro lado las directivas no son directamente aplicables y no a todos los estados, pero, en el caso de que aplique, la legislación española tiene que, dentro de un plazo marcado, incorporar estas directivas en la ley. En materia de protección de la propiedad intelectual existen múltiples directivas y reglamentos, algunos de los cuales aplicables a este proyecto son:

- Directiva (UE) 2019/790 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital y por la que se modifican las Directivas 96/9/CE y 2001/29/CE (Texto pertinente a efectos del EEE.). [4]
- DIRECTIVA 2004/48/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 29 de abril de 2004 relativa al respeto de los derechos de propiedad intelectual (Texto pertinente a efectos del EEE). [5]
- Directiva 2011/77/UE del Parlamento Europeo y del Consejo, de 27 de septiembre de 2011 , por la que se modifica la Directiva 2006/116/CE relativa al plazo de protección del derecho de autor y de determinados derechos afines. [6]

Estas directivas, aunque no directamente aplicables ya que no forman parte directa de la legislación española, establecen una base legal que las instituciones de los países miembros deben adoptar en su marco regulador en un plazo determinado de tiempo. Es por esto por lo que hay que actuar con estas directivas en mente ya que el eventual efecto legal de estas es inevitable.

1.6. Estructura del documento

- **Estado de la cuestión:** Esta sección aborda la situación actual del campo de estudio y presenta el contexto teórico en el que se enmarca el proyecto. Aquí se analizan y presentan las investigaciones previas y las teorías existentes relevantes al tema.
- **Solución propuesta/método:** En esta parte del trabajo se detalla la metodología y enfoque utilizado para abordar la pregunta de investigación y resolver el problema planteado. Se describe cómo se recopilaron y analizaron los datos, qué técnicas y herramientas se utilizaron y por qué se eligió esta metodología en particular.

- **Validación y pruebas:** En esta sección, se detallan los procedimientos utilizados para validar la metodología empleada en el desarrollo y se exponen los resultados de las pruebas realizadas. La validación se llevó a cabo con el fin de garantizar la confiabilidad y precisión de los datos obtenidos.
- **Resultados:** En esta sección se presentan los hallazgos y resultados de la investigación. Dichos resultados se presentan de manera clara y concisa, por medio de tablas y gráficos.
- **Gestión del proyecto:** Esta sección se centra en aspectos relacionados con la planificación y ejecución del proyecto de investigación. Incluye las diferentes fases del desarrollo, el presupuesto del mismo e incluye la sección del impacto socio-económico del proyecto.
- **Conclusiones:** En el apartado de conclusiones se presenta un resumen detallado del desarrollo y los resultados alcanzados en la investigación, se discuten las limitaciones encontradas durante el estudio y se proponen posibles direcciones para futuras investigaciones en el tema.

2. ESTADO DE LA CUESTIÓN

2.1. Objetivos de Desarrollo Sostenible

En 2015 la Organización de las Naciones Unidas (ONU) definió una serie de objetivos con el fin de enfrentar y superar los principales problemas que afectan a la humanidad hoy en día en un plazo de 15 años. Estos se conocen como Objetivos de Desarrollo Sostenible (ODS) y conforman la denominada Agenda 2030, la cual detalla el plan de acción a seguir para guiar el desarrollo humano en una dirección más sostenible [7].

En total fueron definidos 17 objetivos, cada uno desglosado en diferentes metas a alcanzar. Estas metas establecen las bases sobre las cuales los gobiernos, entidades privadas y organismos internacionales deben construir sus políticas con el fin de cumplir dichos objetivos. La colaboración entre las entidades mencionadas es esencial para el éxito de la Agenda, tanto es así, que el último de todos los ODS especifica de forma explícita la importancia de esta cooperación [8].

De manera adicional, la Agenda 2030 detalla una serie de indicadores usados para determinar de forma cuantitativa, objetiva y estandarizada el grado de desarrollo de un país o institución con respecto a cada uno de los objetivos [9]. Estos indicadores, junto con otras medidas proporcionadas por análisis de expertos o soluciones basadas en Inteligencia Artificial (IA) permiten realizar un seguimiento efectivo de los objetivos. Mediante estas técnicas es posible verificar si las políticas adoptadas por las entidades son efectivas a la hora de alcanzar los objetivos o no.

Por ello, el desarrollo de herramientas e indicadores capaces de determinar de forma objetiva la validez de las medidas adoptadas es de vital importancia. Además, en el caso de no alcanzarse alguno de los objetivos, toda la información obtenida mediante estos análisis permitirá delimitar los factores que causaron el incumplimiento de la Agenda 2030. Anualmente, las Naciones Unidas publican en un informe las investigaciones realizadas así como el análisis resultante de la aplicación de los indicadores mencionados anteriormente.[10]

Por su naturaleza, estos objetivos presentan una alta interconexión entre sí, dando lugar a casos en los que avances positivos hacia un objetivo pueden resultar en un impacto negativo hacia otro objetivo diferente. Esta interconexión es un ámbito muy amplio de estudio y de suma importancia para lograr entender el ecosistema de metas relacionadas que conforman los ODS con el fin de poder proporcionar soluciones equilibradas que remen en una dirección común.

Este trabajo, aun no teniendo como objetivo explorar estas interrelaciones, puede emplearse para identificar estas conexiones subyacentes entre diferentes objetivos, facilitando la elaboración de nuevos trabajos de investigación sobre la materia.

A continuación se presentan cada uno de los 17 ODS, aportando información sobre los mismos [9]:

1. **Fin de la pobreza:** Erradicar de manera efectiva la pobreza extrema y la desigualdad económica.
2. **Hambre cero:** Garantizar y asegurar una nutrición adecuada, promoviendo una agricultura sostenible
3. **Salud y bienestar:** Mejorar la salud de la gente, garantizando el acceso a un servicio sanitario competente
4. **Educación de calidad:** Asegurar una educación de calidad, inclusiva, y objetiva para todo el mundo.
5. **Igualdad de género:** Erradicar la desigualdad de género, empoderando a las mujeres y niñas.
6. **Agua limpia y saneamiento:** Realizar un consumo sostenible del agua y garantizar el acceso a agua limpia.
7. **Energía asequible y no contaminante:** Generar energía de manera verde y sostenible, garantizando el acceso a la misma.
8. **Trabajo decente y crecimiento económico:** Garantizar un trabajo decente para todos y fomentar el crecimiento sostenible.
9. **Industria, innovación e infraestructura:** Formar una industria, infraestructura y corrientes de investigación sostenibles
10. **Reducción de las desigualdades:** Reducir las desigualdades entre países y dentro de los mismos.
11. **Ciudades y comunidades sostenibles:** Lograr que las ciudades y los centros de población sean sostenibles.
12. **Producción y consumo responsables:** Formar una cadena de producción sostenible, haciendo un uso eficiente de los recursos.
13. **Acción por el clima:** Frenar y combatir el cambio climático
14. **Vida submarina:** Conservar y hacer un uso responsable y sostenible de los mares y recursos marinos
15. **Vida de ecosistemas terrestres:** Conservar y hacer un uso responsable y sostenible de los ecosistemas terrestres
16. **Paz, justicia e instituciones sólidas:** Promover sociedades pacíficas y un sistema jurídico justo y accesible para todos.

17. **Alianzas para lograr los objetivos:** Promover la cooperación y la formación de instituciones internacionales para cumplir los objetivos.

2.2. Inteligencia artificial

La Inteligencia Artificial (IA) es la disciplina que compone todas aquellas técnicas que permiten a un ordenador tomar decisiones informadas de manera automática y precisa, imitando en mayor o menor medida las características y funciones cognitivas propias del cerebro y la mente humana [11]. Se trata de un área de conocimiento extensa, en la cual se está invirtiendo una gran cantidad de recursos, tanto en desarrollo como en investigación, y se encuentra a la vanguardia tecnológica. Se trata de un área de investigación relativamente antigua en relación con la computación, llegando a preceder a los ordenadores como los conocemos hoy en día.

Dentro del estudio de la IA existen diferentes disciplinas de gran relevancia actualmente como lo son el Aprendizaje Automático y el Aprendizaje Profundo. El primero de ellos hace referencia a aquellas inteligencias artificiales que son capaces de aprender con el tiempo basándose en unos datos de ejemplo. Aprendizaje Profundo es a su vez un subconjunto de Aprendizaje Automático en el que las máquinas están formadas por una serie de capas de neuronas interconectadas, de manera análoga a la biología del cerebro humano [12].

El Aprendizaje Automático es un campo de estudio muy extenso, sin embargo su objetivo radica en la generación de modelos matemáticos capaces de identificar patrones en un conjunto de datos mediante un proceso de entrenamiento o aprendizaje. Posteriormente, el modelo entrenado se emplea para proporcionar una solución a un problema vinculado a dichos datos. Dentro del Aprendizaje Automático se pueden hacer numerosas subdivisiones en diferentes disciplinas en función de diversos criterios. Sin embargo, para este trabajo, el criterio que más procede es el que se realiza en base a la naturaleza de los datos que se emplean como entrenamiento. Siguiendo por esta línea, los diferentes conjuntos que conforman el campo del Aprendizaje Automático son:

- **Aprendizaje supervisado:** Estos modelos hacen uso de datos *en bruto*, sin ningún tipo de información adicional, llamados no etiquetados. Estos modelos necesitan realizar un análisis profundo para identificar las características que diferencian a los datos entre si. Existen diferentes aproximaciones que usar al la hora de tratar con datos no etiquetados:
 - Modelos probabilísticos: usan funciones probabilísticas para inferir relaciones y características nuevas de los datos dados
 - Modelos de agrupación: hacen uso de medidas de distancia para agrupar los datos en grupos independientes, revelando relaciones ocultas entre ellos.

- **Aprendizaje no supervisado:** Suelen ser los más potentes pero necesitan de datos etiquetados, datos que ya contienen la información que se desea extraer de los mismos, de esta manera al presentarles datos nuevos serán capaces de inferir esta información deseada en base a los datos dados para el entrenamiento. Algunos son los siguientes
 - Árboles de decisión: modelos en forma de árbol donde las salidas deseadas se encuentran en las hojas y un dato dado va iterando por cada rama eligiendo un camino u otro dependiendo de los atributos diferenciadores mas importantes hasta generar la salida deseada. Estos atributos y los valores que se usan como baremo para las ramificaciones son sacados de los datos de entrenamiento.
 - Aprendizaje Profundo: probablemente el tipo de modelo más importante y que más en auge está. Su objetivo es imitar el comportamiento del cerebro humano, simulando las neuronas y sus conexiones para generar, en base al conocimiento almacenado en forma de datos, un agente inteligente. Tiene un problema importante que es la falta de interpretabilidad de los modelos, estos en la mayoría de los casos funcionan como una caja negra a la que se le introducen datos y genera salidas, pero el funcionamiento interno y el por qué del mismo muchas veces es incomprensible.

2.2.1. Historia de la IA

El surgimiento de la IA suele enmarcarse en la década de los 40, cuando el matemático Alan Turing conocido como el padre de la computación moderna planteó el famoso *Test de Turing* [13]. En él se plantea si una máquina pudiera llegar a pensar de forma autónoma e imitar el comportamiento humano. Este test consiste en una intervención entre dos humanos y una máquina. Uno de los dos humanos realiza preguntas de manera simultánea a los dos participantes restantes, el otro humano y la máquina. Este tiene como objetivo determinar cual de las dos respuestas ha sido generada por la máquina. Por otra parte, la máquina tiene como objetivo hacerse pasar por humano, por lo tanto, si no es reconocida como máquina superará la prueba.

Fuera del marco teórico, la primera evidencia experimental que hay sobre la IA tal y como la conocemos en la actualidad es la creación de la célula de McCulloch Pitts Fig. 2.1. Estas células, o dispositivos, tienen dos estados posibles apagado (0) y encendido (1) y cuentan con una serie de conexiones de entrada y una de salida, estas conexiones alteran el valor que pasa por ellas acorde al peso de dicha conexión, de esta manera simulan una versión simplificada de una neurona humana. Estas neuronas, aunque simples suponen la base para el resto de modelos de Aprendizaje Automático.

El siguiente gran paso en la IA fue la creación del perceptrón multicapa, basado en la interconexión de Células de McCulloch Pitts. Su creación supuso un gran paso puesto que introdujo unos de los conceptos más importantes en el desarrollo de Redes de Neuronas

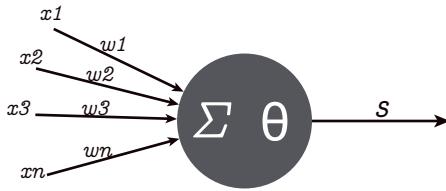


Fig. 2.1. Modelo célula McCulloch Pitts

Artificiales, *la Retropropagación*. En todos los modelos basados en neuronas (Aprendizaje Profundo) el peso de las conexiones necesita ajustarse a los datos de entrada, es esto lo que las dota de la capacidad de aprender. En una red con una sola capa es trivial, primero se calcula el error cometido en la fase de entrenamiento y posteriormente se ajustan las conexiones acorde a ese error. Pero con más de una capa de neuronas, el problema se vuelve más complejo puesto que no se sabe en qué medida hay que ajustar las conexiones de cada capa. Es este problema el que se soluciona con el algoritmo de retropropagación, haciendo uso de funciones de error derivables para inferir el error proveniente de cada capa. Este descubrimiento abrió las puertas a la creación de arquitecturas neuronales cada vez más complejas y potentes, encontrándose actualmente la limitación de estos modelos en su complejidad computacional y la cantidad y calidad de los datos que se emplean para el entrenamiento.

Después de la creación del perceptrón y el algoritmo de retropropagación, la investigación en el ámbito de la IA atravesó un periodo de altibajos, los llamados inviernos de la IA. Épocas en las que la investigación se ve detenida bruscamente debido a barreras tecnológicas y, a su vez épocas doradas que surgen a raíz de descubrimientos punteros que rompen esas barreras. Hoy en día gracias a los avances en capacidad computacional y la gran cantidad de datos disponibles nos encontramos en la época más fructífera de la IA. Esto está apoyado por la creación de una serie de modelos basados en aprendizaje profundo altamente especializados en tareas concretas. Las características y arquitecturas de estos modelos están en gran medida influenciadas y determinadas por la naturaleza de los problemas que se espera que resuelvan. Algunos de estos modelos importantes son:

Redes neuronales prealimentadas

Conforman la familia de redes neuronales artificiales más sencillas, definida por todas las redes de neuronas en las que la entrada se propaga en una sola dirección, hacia adelante. Pese a que su definición hace que sea un término aplicable a multitud de redes, es común utilizarlo a la hora de hacer referencia a un perceptrón multicapa.

Un perceptrón, como se ha mencionado anteriormente, es una de las primeras redes diseñadas y la primera que podemos considerar compleja. Su arquitectura es muy básica, consta de una serie de capas con un número arbitrario de neuronas en cada una de ellas. Todas las neuronas de una capa están conectadas a todas las neuronas de siguiente. Es gracias a esta característica por la que estas redes también son denominadas como *Fully*

Connected.

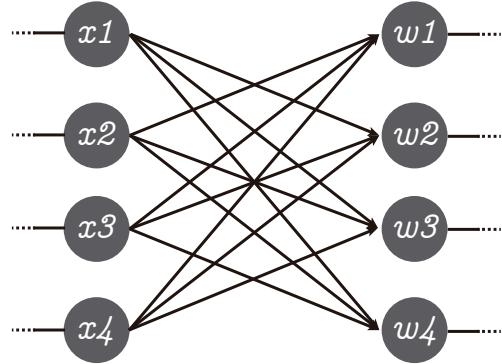


Fig. 2.2. Representación perceptrón multicapa

Redes convolucionales

Tienen como objetivo principal el procesar y tratar imágenes, es por esto por lo que la red está formada por una serie de capas en dos dimensiones, representando la imagen. Estas capas hacen uso de convoluciones para extraer los aspectos y características generales más importantes de las imágenes dadas, es por esto por lo que las denominamos como *capas de convolución* o *convolucionales*. Estas extraen la información reduciendo el tamaño de la entrada cada vez más y eliminando las partes menos relevantes. Posteriormente a estas capas convolucionales, la señal de la red, se ve *aplanada*, pasando de dos a una dimensión para poder servir como entrada a un red prealimentada. Es esta última red la que procesa esta información generando la salida deseada.

Estas redes han demostrado ser increíblemente potentes y versátiles a la hora de tratar con imágenes ya que las capas de convolución, al extraer la información más relevante de cada imagen hace que identificar los patrones necesarios para resolver el problema deseado sea una tarea más sencilla.

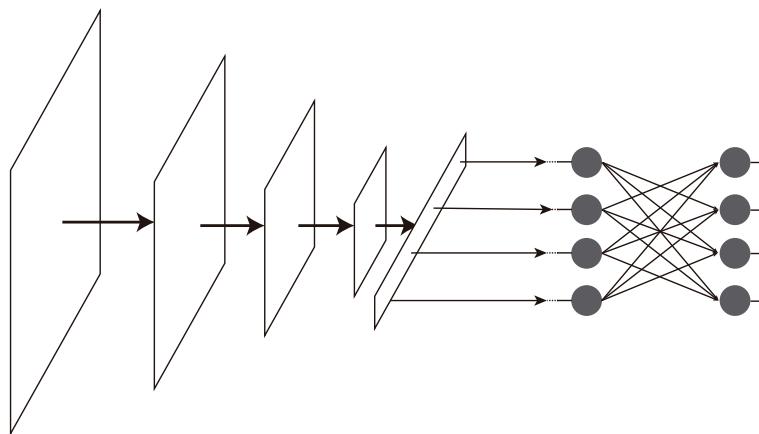


Fig. 2.3. Representación red convolucional

Redes recurrentes

Son el primer tipo de redes que se salen de la definición de *prealimentada*. Estas hacen usos de mecanismos que implementan características similares a la memoria, incorporando como entrada a la red la salida generada anteriormente, esto proporciona a la red datos contextuales, variando la salida para una misma entrada dependiendo de este contexto.

Estas características hacen de las redes recurrentes excepcionales modelos a la hora de tratar con datos secuenciales, es decir cuya información no solo esté codificada en el dato en si, si no que esta dependa contexto que lo rodea. Se usan especialmente en el ámbito del procesamiento del lenguaje natural, puesto que el significado de una palabra varía dependiendo del contexto en el que se encuentra.

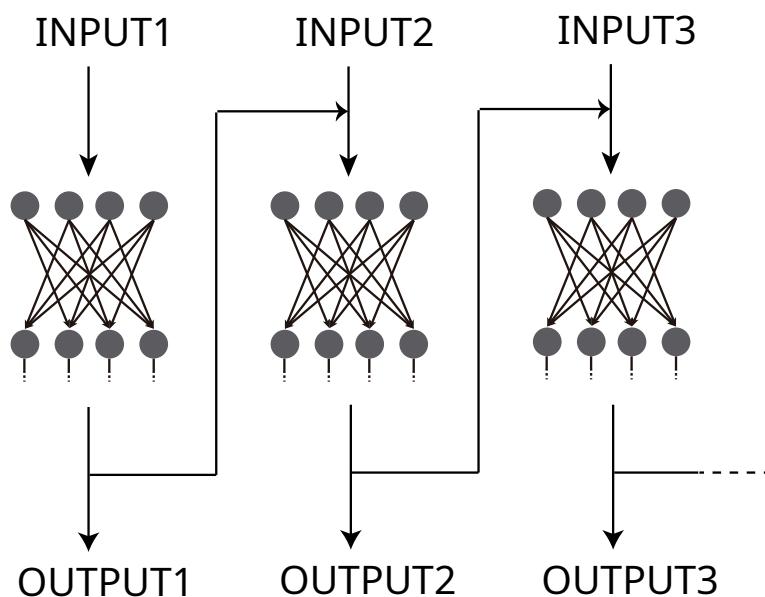


Fig. 2.4. Representación red recurrente

Mecanismos de atención

En los últimos años el término *Mecanismos de atención* ha cobrado una notable importancia en el ámbito del aprendizaje profundo. Hace referencia a la aptitud humana de centrarse , o en mayor medida, en partes concretas de las entradas (los sentidos), idealmente en las partes más relevantes. Esto es especialmente importante ya que reduce notablemente la capacidad de procesamiento necesaria para interpretar la información captada. Este concepto aplica de manera directa a las redes de neuronas, haciendo que si estas implementan mecanismos de atención, sean capaces de aprender en menos tiempo, utilizando menos recursos y posiblemente ignorando el ruido que puedan contener las entradas.

La atención en el aprendizaje profundo se suele implementar como una matriz en la que se relacionan, contextualmente, todos los elementos de la entrada entre si. De esta manera, al procesar uno de estos elementos se puede identificar la importancia que este tiene con respecto a todos los demás y prestar **atención** a aquellos que están más relacio-

nados. Esto difiere con los modelos de redes recurrentes, ya que estos solo relacionan una entrada con la anterior, implementando mecanismos de atención se puede tener una foto de las relaciones entre todos los elementos.

Las ventajas que presentan estos mecanismos se ve aparente cuando se quieren procesar frases como la siguiente: “El pangolín dormía plácidamente colgado de la rama de un árbol usando su cola.” [14], en este caso una red recurrente, al procesar la última parte de la frase “usando su cola” no podrá identificar a quien se hace referencia. Por otra parte, un mecanismo de atención si podrá identificar la relación entre “El pangolín” y “usando su cola”, pudiendo entender correctamente la frase.

Estos mecanismos se utilizan a la par de redes recurrentes, dotándolas de características de atención, haciendo de ellas unos modelos mucho más completos. Es aquí donde surge uno de los avances más grandes de los últimos años en el ámbito del aprendizaje profundo. Este surge en el 2017 con la publicación “*Attention is all you need*“ [15] y supone un cambio en el paradigma ya que lo autores de dicho artículo proponen una nueva arquitectura que incorpora únicamente estos mecanismos de atención y prescinden de las redes recurrentes, el llamado *transformer*. Estos nuevos modelos, pese a incorporar únicamente mecanismos de atención, presentan una arquitectura compleja, constando de diferentes módulos que, conjuntamente, identifican la información relevante de la entrada y posteriormente, haciendo uso de redes prealimentadas (al igual que con las redes convolucionales), procesan esta información extraída por los mecanismos de atención y generan la salida deseada.

2.2.2. Aprendizaje por transferencia

Unos de los conceptos más importantes en la actualidad es el llamado *Transfer Learning* (Aprendizaje por transferencia), este hace referencia a utilizar un modelo entrenado con un conjunto de datos inicial y que ha obtenido buenos resultados y entrenarlo con un conjunto nuevo. Habitualmente ese conjunto inicial está formado por datos generales para establecer una base de conocimiento amplia para posteriormente adaptarla a un problema nuevo, de manera más rápida y eficaz que si se entrenara únicamente con datos concretos, afines al problema a resolver.

Hoy en día tiene especial importancia ya que los modelos que mejores resultados obtienen cuentan con arquitecturas complicadas y muy pesadas, los mencionados transformers, estando formados por cientos de millones de parámetros variables. Estos modelos solo pueden ser entrenados con conjuntos de datos masivos y en ordenadores con una alta capacidad computacional y una amplia memoria disponible. Es por esto por lo que no cualquiera puede generar y menos entrenar un modelo de estas magnitudes por lo que entidades grandes los desarrollan para que posteriormente puedan ser usados en un corto periodo de tiempo y haciendo uso de conjuntos de datos menores, adaptados a las necesidades finales.

Existen multitud de modelos disponibles con los que hacer aprendizaje por transferencia dependiendo de la finalidad deseada, de todos ellos el más relevante en el marco de este proyecto es:

BERT (Bidirectional Encoder Representations from Transformers) Modelo publicado en 2019 por Google es uno de los más utilizados para tareas relacionadas con el tratamiento de texto. Como su nombre indica consta de una arquitectura basada en transformers, pero cuenta con un factor diferenciador ya que presenta una capacidad de atención bidireccional, esto significa que las palabras se relacionan con las anteriores y posteriores.[16] La potencia del modelo surge de esta arquitectura y la manera en la que se ha entrenado, este entrenamiento también es peculiar, contando con dos fases, cada una con el objetivo de perfeccionar capacidades diferentes, estos dos entrenamientos son:

- Masked LM: en este primer proceso de entrenamiento se ocultan ciertas palabras de la entrada de manera aleatoria, al rededor de un 15 %, y se entrena el modelo para que sea capaz de predecir cuales son estas palabras, basándose únicamente en el contexto que las rodea. El proceso de ocultado también es relevante ya que los elementos seleccionados no siempre se reemplazan con una cadena específica que indica esta ocultación, si no que otras veces se reemplazan por otro elemento aleatorio y otras veces no se cambian pese a ser seleccionados. Esta parte proporciona un entrenamiento bidireccional.[16]
- Next Sentence Prediction (NSP) "Predicción de la Siguiente Frase": en esta segunda fase se entrena el modelo para que sea capaz de predecir la siguiente frase, como los desarrolladores afirman, esta parte fortalece las capacidades de respuesta a preguntas y de inferencia del lenguaje natural del modelo. Para este entrenamiento se escogen dos frases consecutivas cualesquiera *A* y *B* y se entrena a el modelo para que dado *A* sea capaz de predecir *B*. Este paso también cuenta con peculiaridades, ya que no todas las veces la frase *B* seleccionada es realmente consecutiva a *A*, concretamente en el 50 % de los casos es una frase aleatoria del corpus.[16]

Estos dos tipos de entrenamientos hacen de BERT un modelo sorprendentemente versátil, robusto y sorprendentemente fácil de usar, siendo necesario añadir una única capa adicional para poder adaptarlo a virtualmente cualquier aplicación de NLP, contenido esta última capa las neuronas necesarias para codificar la salida deseada.

Otros modelos de *aprendizaje por transferencia* comunes son:

- **GPT (Generative Pre-trained Transformer)**: Los modelos GPT son arquitecturas Transformer preentrenadas que se utilizan para generar texto coherente y contextualmente relevante. Estos modelos son entrenados en grandes cantidades de texto y son capaces de generar texto en función de un contexto dado.

- **ResNet:** ResNet es una arquitectura de red neuronal convolucional que utiliza bloques residuales para permitir un entrenamiento efectivo de redes profundas. Esto es especialmente útil en tareas de visión por computadora donde se deben identificar objetos en imágenes.
- **InceptionNet:** InceptionNet, también conocida como GoogLeNet, es una arquitectura de red neuronal convolucional que utiliza módulos de convolución múltiple para capturar características de diferentes escalas en una imagen.
- **VGGNet:** VGGNet es una arquitectura de red neuronal convolucional que se destaca por su simplicidad y profundidad uniforme. Es conocida por su rendimiento en tareas de reconocimiento de objetos en imágenes.
- **Xception:** Xception es una arquitectura de red neuronal convolucional que se basa en la idea de separar las convoluciones espaciales y de profundidad. Esto ayuda a capturar detalles y características en imágenes de manera eficiente.
- **MobileNet:** MobileNet es una familia de arquitecturas de redes neuronales convolucionales diseñadas para ejecutarse eficientemente en dispositivos móviles y embedidos, lo que las hace ideales para aplicaciones con recursos limitados.
- **DenseNet:** DenseNet es una arquitectura de red neuronal convolucional que utiliza conexiones densas entre capas para permitir un flujo eficiente de información a través de la red. Esto ayuda en la identificación precisa de objetos en imágenes.
- **U-Net:** U-Net es una arquitectura de red neuronal utilizada en tareas de segmentación de imágenes, especialmente en aplicaciones médicas. Su estructura de codificador-decodificador es efectiva para resaltar áreas específicas en imágenes.
- **CycleGAN:** CycleGAN es un tipo de red generativa adversaria (GAN) que se utiliza para transformar imágenes de un dominio a otro, como convertir fotos en pinturas al estilo de un artista o cambiar el estilo de una imagen.

2.3. Procesamiento del Lenguaje Natural

Dentro de la IA todo el subgénero que se dedica a tratar con texto se denomina *Procesamiento del Lenguaje Natural* o NLP por sus siglas en inglés. Esta disciplina tiene como objetivo general abrir un canal de comunicación entre las máquinas y los humanos, manteniendo la independencia de las partes comunicadoras. Para esto un ordenador tiene que ser capaz de entender el lenguaje como lo haría un humano, tanto para el procesamiento de las entradas textuales que le llegan como para, a partir de esas u otras entradas de otros tipos, generar un texto comprensible por un humano. Una vez un ordenador es capaz de entender texto natural se puede entrenar para realizar tareas como la traducción, clasificación y generación coherente de texto. [17]

Se ha mencionado el *Lenguaje Natural*, pero es importante destacar la diferencia entre este tipo de lenguajes y los formales, tales como el lógico o matemático. Este primero surge de la necesidad humana de comunicarse, y evoluciona como lo haría un organismo vivo para conseguir una comunicación lo más completa y eficiente posible y cuenta con una serie de reglas que lo definen pero no lo limitan y estas evolucionan en paralelo. Esto difiere de los lenguajes formales en el hecho de que estos últimos tienen una serie de reglas estrictas e invariantes que definen el uso que ha de hacerse del lenguaje y cualquier variación o excepción no es aceptada.[\[18\]](#)

Una vez definido el lenguaje natural solo queda interpretarlo, esto es un problema ya que su representación no es matemática, por lo que un ordenador, y en concreto y modelo de inteligencia artificial no sería capaz de procesarlo. Es por esto que el primer paso es seguir algún proceso para representar de manera matemática las entradas textuales. Este preprocesado requiere de varios pasos en lo que se divide la entrada en diferentes partes o dosis de información llamados *tokens*, también se requiere de normalizar dichas palabras, eliminando información irrelevante y dejando únicamente la raíz, puesto que es la parte que contiene el mayor número de información. Una vez procesada la entrada se traduce a su representación matemática para poder ser tratada por un ordenador.

Tokenización

La primera de estas tareas es la tokenización, esta consiste en dividir la entrada de texto en sus partes fundamentales. A priori parece una tarea sencilla, dividir la entrada en sus diferentes palabras, realizando una división en cada espacio. Esto, aunque intuitivo no siempre funciona, sobre todo en inglés ya que muchas palabras se representan con apóstrofes y tienen que ser divididas. Es por esto que se hace uso de una serie de reglas y expresiones regulares que tienen en cuenta todas estas posibilidades y consiguen dividir de manera eficaz y eficiente el texto en una serie de *tokens*.

The house wasn't sold for \$ 500,000.00

The_house_was_n't_sold_for_\$.500.000,00

Lematización

El siguiente paso después de dividir la entrada es normalizarla, este proceso consiste en estandarizar las entradas para que todas contengan información del mismo tipo y estructura [\[19\]](#). En el uso de números la normalización consiste en transformar todas las entradas para que estas estén en un rango entre 0 y 1, manteniendo las relaciones pero normalizando la magnitud. En el ámbito de textos la normalización suele consistir en reducir las palabras a su raíz, la llamada *lematización*, haciendo que todas mantengan la mayor parte de su significado pero eliminando las diferencias irrelevantes entre ellas y

haciendo que dos palabras diferentes con el mismo significado sean virtualmente iguales (*soy = eres*). [20]

The house wasn't sold for \$ 500,000.00

The_house_be_not_sell_for_\$_500.000,00

Vectorización

El último paso es traducir la entrada ya preparada para que pueda ser utilizada por un ordenador, esto se hace transformando cada palabra en un vector que la representa y se denomina *Word Embedding*. Esto se puede llevar a cabo por medio de varios métodos pero actualmente uno de los más usados es el llamado *word2vec*[21], [22]. Este no es un método como tal, si no dos técnicas usadas para entrenar modelos de aprendizaje automático con la finalidad de vectorizar palabras, de tal manera que palabras similares se encuentran en posiciones similares. Estas dos técnicas son *Continuous Bag of Words (CBOW)* y *Continuous Skip-gram (SG)*, ambas definen métodos de entrenamiento para que los modelos entrenados siguiendo esas pautas sean capaces de vectorizar tokens correctamente, ambos tienen como objetivo predecir palabras pero de maneras diferentes.

- CBOW: en esta arquitectura se usan las palabras anteriores y posteriores (4 en cada dirección suele generar los mejores resultados) para intentar predecir la palabra del medio.
- SG: esta arquitectura funciona de manera inversa a la anterior, en esta se usa una palabra para intentar predecir ciertas palabras anteriores y ciertas posteriores

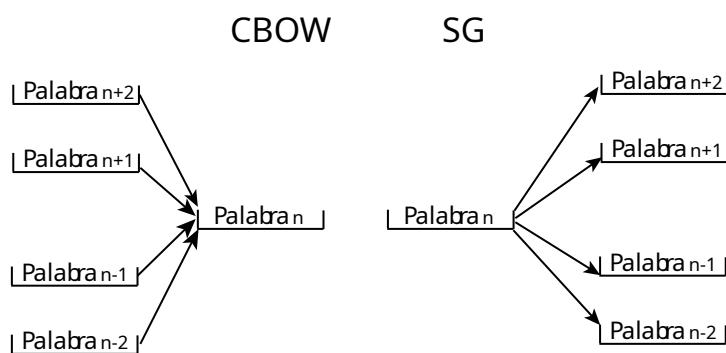


Fig. 2.5. CBOW vs SG

Estas dos aproximaciones pueden implementarse por medio de varios modelos, en el trabajo original los autores usan un clasificador *log-lineal*, modelo que hace uso de funciones para, a partir de las características de los datos, generar distribuciones probabilísticas que posteriormente son usadas para tareas de clasificación y predicción[21]. El hecho de trabajar en escala logarítmica los hace especialmente eficaces a la hora de identificar y

modelar relaciones complejas entre los datos [23]. Una vez se ha entrenado el modelo, este puede ser usado para generar un vector de la dimensionalidad deseada que encapsula la información semántica de la palabra, esta información surge a raíz del aprendizaje realizado en la tarea de predicción.

El hecho de representar las palabras mediante vectores abre un amplio abanico de herramientas de análisis, pudiendo hacer uso de operaciones algebraicas con palabras. Estas operaciones nos permiten, entre otras cosas, calcular la similitud entre palabras, partiendo de la base de que palabras similares tienen vectores asociados que se encuentran en posiciones similares en el espacio, si calculamos el producto escalar entre los vectores de dos palabras obtenemos como resultado una medida de similitud entre las dos. Adicionalmente el hecho de definir un espacio de alta dimensionalidad permite el cálculo de diferentes medidas de similitud, dependiendo de las dimensiones y operaciones usadas.

Otra característica de estos vectores es que pueden ser sumados y restados, resultando en propiedades curiosas como por ejemplo la siguiente operación: $v[\text{rey}] - v[\text{hombre}] + v[\text{mujer}]$. El resultado de esta operación es un vector que se encuentra cerca de $v[\text{reina}]$.

Una vez se ha procesado y vectorizado el texto a usar, se pueden usar los modelos y técnicas definidos para realizar las tareas y solucionar los problemas deseados.

2.4. Clasificación multi etiqueta

Una de las tareas más comunes es la de la clasificación, esta consiste en que el modelo sea capaz de predecir a que clase pertenece un dato dado, algunos ejemplos comunes son la clasificación de números y letras escritas a mano y la clasificación de textos acorde al sentimiento que evoca. Estas tareas se pueden clasificar en dos grupos en base al número de clases predecidas y los datos usados:

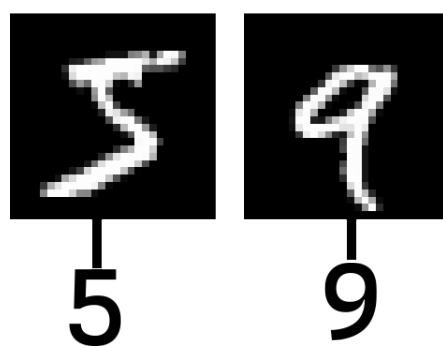


Fig. 2.6. Clasificación números

Clasificación multiclasa En este caso se desea clasificar a los datos en una de varias clases, de manera que cada dato pertenece únicamente a una a la vez. Por ejemplo el caso mencionado anteriormente de la clasificación de números, figura 2.6, una única imagen

puede corresponder a un único numero.

Clasificación multietiqueta En este caso el objetivo es clasificar cada dato en todas las clases a la que pertenezca, pudiendo un dato pertenecer a varias a la vez, por ejemplo el caso de clasificación de textos de acorde al sentimiento que evocan, un texto puede tener un tono enfadado y agresivo a la vez. [24]

Cabe mencionar un último tipo de clasificación, la llamada *clasificación binaria*, en este tipo de tareas se quiere asignar uno de dos valores a una entrada, positivo o negativo por lo que solo se cuenta con una clase.

2.4.1. Función de activación

Una de las partes esenciales en las neuronas de una red es la función de activación, esta transforma el valor que se propaga por la red. En las neuronas de salida tienen un papel fundamental siendo las responsables de generar una salida con el formato correcto e interpretable.[25]

En el caso de clasificación esta función de activación final puede ser de dos tipos dependiendo del problema de clasificación *sigmoid* y *softmax*.[26]

Sigmoid Una de las funciones de activación más comunes ya que se trata de una función no lineal, lo cual permite a las redes adaptarse a problemas no lineales, y genera salidas entre 0 y 1. Estas salidas, en tareas de clasificación, pueden ser tratadas como probabilidades por lo que la salida se puede interpretar fácilmente, asignando a una entrada todas aquellas clases cuya respectiva salida supera un determinado umbral. Este proceso aplica de igual manera a las tareas de clasificación binaria, para las cuales se debe contar con una única neurona de salida. [27]

Softmax Esta función de activación está formada por una combinación de funciones sigmoid que, al igual que esta, genera un vector con una probabilidad por cada neurona de salida y por ende cada clase posible. La principal diferencia entre las dos es que en este caso todas las probabilidades suman 100 % por lo que no son independientes. Esto hace de softmax la mejor función de activación para tareas de clasificación multiclas, cuando se quiere asignar una única etiqueta a la entrada, en este caso se asigna aquella etiqueta con la mayor probabilidad asignada. [25]. Los vectores usados para entrenar este tipo de redes se denominan como *one-hot*², teniendo un único elemento con un valor positivo de 1, mientras que el resto tiene un 0. El proceso de clasificación transforma el vector de salida en uno de este tipo, asignando un 1 a la clase con mayor porcentaje y un 0 al resto.

²La codificación one-hot es una técnica de representación de datos en la que cada valor categórico se convierte en un vector binario único, se trata como un anglicismo debido a la falta de una traducción directa en español.

2.5. Web Scraping

Una vez definido el estado del arte referente a la Inteligencia Artificial, es fácil deducir que obtener una buena cantidad de buenos datos es imperativo si se desea obtener un buen modelo, para esto una de las técnicas más usada es la de *Web Scrapping* o *raspado de web* según su traducción literal. Esta es una técnica en la que se desarrollan agentes automáticos que navegan por internet con la finalidad de extraer los datos deseados de uno o varios sitios web. Hay muchas maneras de conseguir este objetivo, pero todas se engloban en dos grupos, los que imitan las acciones que haría un humano, interactuando de manera dinámica con una página por medio de un navegador y los que hacen uso de peticiones HTTP simples y sus respectivas respuestas para inspeccionar el código fuente. Estos últimos no hacen uso de navegadores por lo que extraen esa información de manera más rápida y eficiente pero corren el riesgo de ser bloqueados por algún sistema de control de tráfico, adicionalmente, si el acceso a los datos requiere de interacción con la página estos últimos agentes no son capaces de recolectarla. [28]

2.6. Tecnologías usadas

En este trabajo se han hecho uso de las siguientes tecnologías y herramientas:

Python Se ha seleccionado python como el lenguaje de programación del proyecto por dos razones principales: actualmente es el lenguaje más utilizado para la recolección y tratado de datos al igual que para realizar trabajos de aprendizaje automático [28]-[30] y es el lenguaje más usado durante la carrera ya que se enseña en primero debido a su sencillez, potencia y versatilidad.

Librerías Las diferentes librerías desarrolladas para python usadas en el desarrollo de este trabajo son las siguientes:

- **Tensorflow:** una de las librerías mas utilizadas para realizar tareas de aprendizaje automático, cuenta con multitud de herramientas para el desarrollo de modelos de aprendizaje profundo, al igual que facilitar el aprendizaje por transferencia, ofreciendo multitud de modelos de manera sencilla.
- **Keras:** librería encapsulada en tensorflow pero de suma importancia ya que es la que implementa todas herramientas de alto nivel para construir y entrenar los modelos, al igual que para tratar los datos a usar en el entrenamiento.
- **Pandas:** Librería por defecto a la hora de tratar con grandes cantidades de datos, permitiendo el uso de estructuras llamadas *DataFrames* que incorporan multitud de herramientas para el procesado y tratado eficiente de datos.

- **Scikit-Learn:** Librería ampliamente utilizada en el ámbito del aprendizaje automático debido a la multitud de herramientas que proporciona
- **Selenium:** Librería más usada a la hora de realizar web scrapping de manera interactiva, por medio de navegadores como *chromium*.
- **Requests y BeautifulSoup4:** Librerías más usada para tareas de WebScraping no interactivo, haciendo uso de requests para realizar las peticiones http junto a BeautifulSoup4 para interpretar las respuestas conteniendo el código fuente HTML de la página facilitando la extracción de datos.
- **Bokeh:** Librería altamente versátil para la creación de gráficas con un alto nivel de personalización y la posibilidad de exportarlas en HTML y javascript para implementarlas de manera interactiva en páginas web.

2.7. Trabajos similares

Durante la fase de investigación se han encontrado multitud de trabajos y herramientas cuya finalidad es similar a la de este trabajo, más en concreto el desarrollo de una herramienta clasificadora que sea capaz de identificar los objetivos de desarrollo sostenible mencionados y presentes en un texto. En esta sección se detallan estos trabajos.

SDG meter Proyecto desarrollado a finales de 2022 por Naciones unidas, en concreto por el Programa de las Naciones Unidas para el Medio Ambiente (UNEP), en el se desarrolla un modelo basado en BERT y una interfaz web que lo encapsula y permite su utilización. En dicha plataforma el usuario puede introducir un texto o archivo y, haciendo uso del modelo, se mostrarán los ODSs relacionados con el texto dado y su respectivo porcentaje de pertenencia. [31]

En su investigación explican las técnicas usadas, las cuales son similares a las descritas en este trabajo ya que realizan un aprendizaje de clasificación multi-etiqueta, por medio del uso de BERT como base para su modelo y *Web Scrapping* para recolectar los datos. Destacar que no describen en detalle el desarrollo llevado a cabo, indicando que usaron el modelo BERT al que se le añadió una capa de salida, por la naturaleza del problema descrito se puede deducir que se trata de una capa con 17 neuronas de salida, y debido a que se trata de un problema de clasificación multi-etiqueta, una capa con función de activación softmax. Esta última afirmación no encaja con los resultados mostrados, si la última capa contara con una función de activación softmax se esperaría que los porcentajes predecidos por el modelo fueran relativamente altos para algunos casos que no se espera que el modelo clasifique, y los casos que el modelo si se espera que clasifique deberían tener asignado un porcentaje alto, idílicamente más de un 50 %. En el caso descrito por los autores asignan predicciones como positivos verdaderos cuando estas tienen un porcentaje de un 1 % dado por el modelo.

En general no explican en mucho detalle la metodología seguida ni la manera en la que interpretan los resultados ni calculan la precisión del modelo, de todas formas afirman haber obtenido un 98 % de precisión.

SDG detector Proyecto publicado por *Journal of Open Source Software (JOSS)* en 2023 y desarrollado por miembros de diversas universidades se trata de una herramienta creada para el lenguaje de programación R, comúnmente usado en el ámbito de la estadística y análisis de datos, que permite identificar a nivel de frases dentro de un corpus textual los ODSs relacionados, junto con las respectivas metas. Siguen una metodología de consultas basadas en palabras clave, similar a las consultas realizadas en bases de datos. [32]

Reportan una precisión de al rededor de un 75 %, teniendo en cuenta que se clasifican las metas y que cuenta con una resolución de clasificación a nivel de frase, es un buen resultado, estando en lo que se suele considerar como el umbral de lo aceptable.

OSDG Plataforma alojada en una web que hace uso de varias técnicas para signar objetivos a un texto, ya sea en formato plano o Portable Document Format (PDF), también cuenta con una Application Programming Interface (API) disponible para clasificar grandes cantidades de datos. Esta herramienta, desarrollada en 2020, en su versión 2.0 cuenta con dos clasificadores independientes, uno basado en una ontología que hace uso de una serie de palabras clave para realizar la tarea de clasificación y uno implementado en esta última versión 2.0 basado en técnicas de Aprendizaje Automático, un objetivo será clasificado por esta herramienta si ambos clasificadores coinciden en su pertenencia. [33]-[35]

Pese a su capacidad de clasificar datos de manera masiva por medio de una API, y puesto que ya se ha usado con esos fines, los autores no proporcionan ninguna métrica de precisión del modelo por lo que su fiabilidad no es conocida.

Linked SDG Producto desarrollado por el Departamento de Asuntos Económicos y Sociales de las Naciones Unidas (UN DESA), consiste en una página web en la que introduce una página web o un archivo del que se extrae una serie de datos relacionados con los objetivos de desarrollo sostenible, las diferentes zonas geográficas presentes en el texto y los ODSs y metas más relevantes en el texto. [36], [37]

text2sdg Paquete para el lenguaje R desarrollado por miembros de la universidad de Basel en septiembre de 2022. Dicho paquete hace uso de diferentes modelos y técnicas existentes basados en consultas para clasificar un texto según los objetivos más relacionados, en estas consultas se busca la presencia de determinados términos, si todos o algunos se cumplen, dependiendo de la estructura de la consulta, se asignarán los objetivos determinados al texto sobre el que esta se ha realizado, similar a la metodología seguida por los autores de SDG detector. [38]-[40]

TFG UPM Trabajo de fin de grado realizado por un estudiante de la Universidad Politécnica de Madrid (UPM) en 2021. En dicho trabajo se usa un modelo Latent Dirichlet Allocation (LDA) para realizar la clasificación. Este modelo proporciona el nivel de relevancia de cada objetivo con respecto al texto proporcionado, no se realiza una clasificación per se asignando de manera categórica objetivos al texto. Finalmente en la fase de pruebas y análisis se proporciona una precisión del modelo de alrededor del 76 %, 9 textos clasificados correctamente como los más relevantes, 4 como relevantes y 4 no identificados como relevantes. Como se destaca en el trabajo las pruebas realizadas son de una magnitud pequeña como para sacar conclusiones sobre el nivel de precisión del modelo pero sirve como un punto de referencia para futuras herramientas. [41]

Clasificación de objetivos con BERT Uno de los trabajos relacionados con la clasificación de textos de acuerdo a los ODSs más completos. Publicado en 2022, en el los autores desarrollan un modelo basado en BERT capaz de predecir los objetivos relacionados con un texto con una precisión del 94 %, medida excelente para una tarea de clasificación con 17 etiquetas. En el realizan un entrenamiento exhaustivo de BERT en una tarea de clasificación multi etiqueta. [42]

Multi label SDG classification Artículo publicado por investigadores de la universidad de Bologna en 2020 tiene como objetivo principal desarrollar una herramienta que pueda agilizar el proceso de análisis de documentos acorde a los ODSs por parte de Naciones Unidas. El modelo desarrollado consiste en una serie de algoritmos de vectorización, Term Frequency-Inverse Document Frequency (TF-IDF, Global Vectors for Word Representation (GloVe) y Universal Sentence Encoders (USE para calcular la similitud entre un texto a clasificar y multitud de textos representativos de los diferentes objetivos, el llamado corpus, esta similitud se calcula combinando la *Similitud coseno*³ de los diferentes algoritmos de vectorización. Una vez calculada la similitud del texto a clasificar con respecto a todos los textos en el corpus hacen uso de un umbral de similitud clasificando aquellos objetivos con una similitud superior al umbral. En su fase de pruebas reportan una precisión de entre el 90 % y 60 % dependiendo del tamaño del conjunto de datos y su naturaleza. [43]

Resulta un trabajo sumamente interesante y versátil ya que, debido a que hace uso de medidas de similitud, el modelo puede ser fácilmente adaptado a nuevas corrientes en la taxonomía de los objetivos simplemente actualizando el corpus.

³La similitud coseno es una medida que evalúa cuán parecidos son dos vectores o documentos, calculando el coseno del ángulo entre ellos.

3. SOLUCIÓN PROPUESTA/MÉTODO

En este capítulo se expone la arquitectura final del sistema propuesto, así como el proceso de desarrollo del mismo en cada una de las partes y como todas estas funcionan en conjunto.

3.1. Arquitectura del sistema

El sistema final consta de una serie de módulos, serán referidos como tal pese a ser más parecidos a herramientas que a módulos, a diferencia de como se esperaría en un producto interconectado.

La finalidad deseada es la creación de un modelo de Aprendizaje Profundo que sea capaz de, dado un texto, asignar los objetivos de desarrollo sostenible relacionados con los temas tratados en el mismo.

En este tipo de tareas se requiere de tres funcionalidades fundamentales, la extracción y análisis de datos, la creación y validación de modelos y una forma de usarlos fácil y eficiente.

Las arquitecturas de estas funcionalidades están representadas en los siguientes diagramas, por motivos de claridad se muestran por separado. Todas ellas funcionan alrededor de la base de datos. Esta hace la función de piedra angular de la arquitectura, por lo que es a través de este elemento por el que se puede ver la interacción entre los diferentes módulos:

3.1.1. Recogida de datos

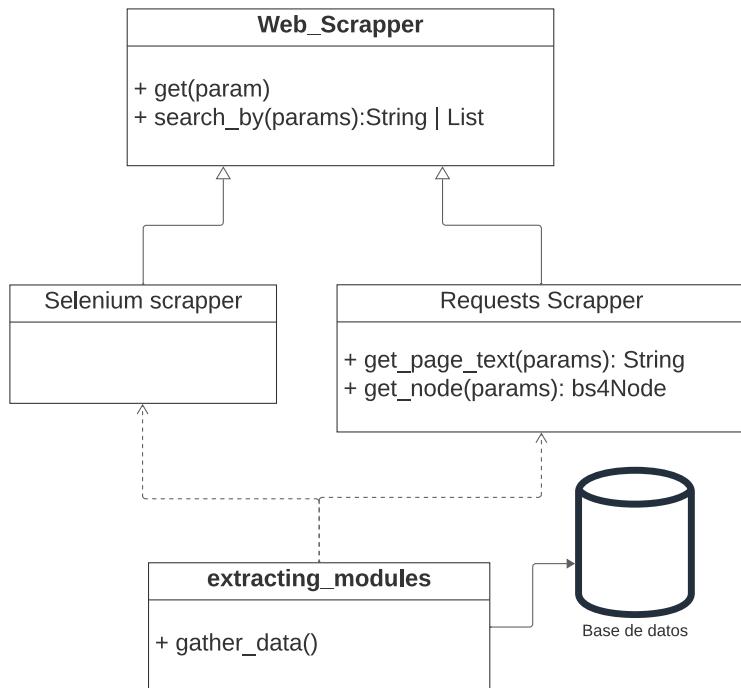


Fig. 3.1. Diagrama UML de la arquitectura de recogida de datos

Para el desarrollo de este sistema se han creado los siguientes módulos:

Módulo de WebScraping

Después de una exhaustiva investigación se llegó a la conclusión de que no existe ninguna base de datos extensa con textos etiquetados relacionados con los Objetivos de Desarrollo Sostenible, es por esto que fue necesario el desarrollo de un módulo de *Web Scrapping*. Este sirve únicamente como herramienta para facilitar la extracción de datos, son otros módulos los que lo usan para implementar la lógica necesaria para realizar la extracción.

La arquitectura de este módulo es sencilla desde fuera pero con múltiples funcionalidades, como se definió en el estado del arte existen dos métodos para realizar esta extracción, es por esto por lo que se creó una clase padre con el esqueleto de una herramienta de este tipo, esta es sencilla constando de dos métodos diferentes:

- Un método *get* que sirve para especificar la página web de la que extraer los datos ya que ambos tienen como primer paso acceder a la página
- Un método *search_by* que sirve para la extracción de los datos.

En la clase padre estos métodos están vacíos, contacto únicamente con un comentario explicando la funcionalidad deseada.

Las dos clases creadas son las siguientes:

requests_scrapper Esta primera clase implementa el tipo de extracción más rápido, la extracción del código fuente devuelto por la petición HTTP directamente, sin ningún tipo de interacción. Como para la extracción de textos habitualmente no se requiere de interacción es, de las dos clases, la más desarrollada.

Implementa un total de cuatro métodos, los dos métodos definidos en la clase padre y dos adicionales, estos son:

- **get:** Método definido en la clase padre, realiza mediante la librería *requests* una petición HTTP a la página especificada, incluyendo en esta una serie de cabeceras para que el destinatario no rechace la conexión. Posteriormente hace uso de la librería *BeautifulSoup4* para interpretar el código HTML y poder extraer datos de la página de una manera sencilla, especificando atributos de los diferentes elementos que conforman la página.
- **get_page_text:** este método tiene una funcionalidad muy simple, únicamente devuelve el texto plano incluido en la página. Gracias a interpretar el código con *BeautifulSoup4* esta tarea es muy fácil ya que el objeto creado por esta librería incluye un atributo el cual contiene el texto de toda la página
- **get_node:** la funcionalidad de este método es parecida a la de get pero en vez de especificar la página en la que buscar se especifica un elemento de la página web. Esto reduce el entorno de búsqueda al contenido de un solo nodo de la página web. Esto es necesario en algunos casos ya que si el nodo que contiene el texto a extraer no tiene ningún parámetro que lo identifique únicamente, esta identificación se puede realizar en un nodo superior que lo contenga, de esta manera es más probable que algún atributo pueda ser usado para identificarlo.
- **search_by:** este es el método más completo y complejo de los cuatro desarrollados ya que realiza la función de extracción de texto. Para la extracción es necesario identificar el nodo HTML que contiene el texto, esto se puede realizar de 5 maneras diferentes: por clase, id, xpath, custom, tag.
 - Por clase: esta es de todas la más útil pero a la vez la más difícil de utilizar ya que una clase puede identificar a varios nodos, de manera que la búsqueda puede resultar en un error o en la extracción del texto equivocado. Esto se soluciona con una serie de filtros. El primero de ellos determina si se desea una extracción múltiple, es decir, el texto de todos los nodos que contengan la clase dada o solo uno, si se quieren todos simplemente de identificar, se extrae el texto y este se devuelve. Si se quiere el texto de uno de ellos es necesario indicar el índice del nodo a extraer, este indica su posición en la lista de todos los nodos que contienen la clase dada, esto se puede hacer de

varias maneras, si se sabe que solo un nodo de la página tiene la clase indicada no se proporciona un índice y se extrae el texto, si este no es el caso pero se sabe que el resto de nodos no contienen ningún tipo de texto se puede extraer el del primer nodo que si tenga y si esto tampoco se puede afirmar habrá que proporcionar la posición deseada.

- Por ID: la más sencilla de todas pero a su vez la menos útil, simplemente se proporciona el ID del nodo objetivo y se extrae el texto de este. Este método no resulta muy útil ya que en la mayoría de los casos los nodos no cuentan con IDs, de hecho se codificó la opción para usarlos pero nunca se desarrolló porque no se encontró ningún nodo con ID.
- Por TAG: TAG es el tipo de nodo de HTML (lo que se encuentra entre <>), esta funcionalidad es útil cuando se reduce el espacio de búsqueda usando `get_node()` ya que en ocasiones se da el caso de que el nodo que contiene el texto no tiene ningún atributo asociado pero alguno de sus nodos padre si y es probable que el tag del nodo deseado si lo identifique únicamente en el espacio de dentro del padre debido a la existencia de nodos específicos para incluir texto.
- Por XPATH: esta funcionalidad se usa como último recurso, cuando el nodo no se puede identificar por ningún otro método. XPATH es la posición absoluta del nodo deseado dentro de la página, del tipo “/html/body/div[1]/div/”, donde se ponen en orden todos los nodos padres separados por “/” y si estos contienen más de un hijo la posición del siguiente. Esta funcionalidad es potente ya que siempre va a identificar a un nodo de manera única pero es altamente susceptible a cambios en el código fuente. De todas formas, si se desea extraer datos de múltiples nodos y estos siguen algún tipo de patrón dentro del código fuente, esta es una funcionalidad extremadamente útil, permitiendo iterar por todos los elementos dentro de un nodo.

Adicionalmente, este método cuenta con dos opciones extra: la opción de filtrar el texto que devuelve, eliminando caracteres especiales, comillas, comas, saltos de línea dobles y codificando el texto en American Standard Code for Information Interchange (American Standard Code for Information Interchange) eliminando todos aquellos caracteres que den error y volviéndolo a codificar en Unicode Transformation Format 8-bit (UTF-8) y la opción de que en vez de devolver el texto se devuelva el nodo seleccionado por si se quiere extraer algún otro tipo de información no textual de este o si se quiere usar para reducir el espacio de búsqueda.

selenium_scrapper Esta segunda clase implementa el otro tipo de extracción, aquella en la cual se interactúa con la página por medio de un navegador para extraer los datos deseados. Esto se realiza haciendo uso de las librerías de *Selenium* para implementar toda la lógica de interacción y extracción y *chromedriver_autoinstaller* para la instalación del

driver web, necesario para permitir que el programa haga uso de un navegador. Esta clase está menos desarrollada que la otra debido a que se ha usado en menos ocasiones por lo que se han identificado menos funcionalidades que implementar, es por esto que cuenta únicamente con los métodos definidos por la clase padre:

- **get**: instancia el driver de chrome en la página web deseada.
- **search_by**: este método es similar al implementado en la otra clase pero con una crucial diferencia, en ningún caso devuelve el texto del nodo si no que devuelve el nodo en sí ya que de esta manera se puede interactuar con la página realizando acciones como hacer click o escribir. De hecho su principal función ha sido extraer las direcciones web de diversos artículos de los cuales posteriormente se ha extraído el texto usando la otra clase.

Módulos extractores de datos

Estos hacen uso de los módulos de *Web Scraping* para extraer los datos, debido a la complejidad de los módulos de scrapeo, este proceso de extracción es sumamente sencillo, teniendo únicamente que inspeccionar las páginas web e indicando que nodos contienen los textos deseados e implementar algún tipo de lógica sencilla como pueden ser bucles para iterar por diversos nodos o interactuar con la página y, finalmente, realizar el almacenamiento de los datos.

3.1.2. Gestión de datos

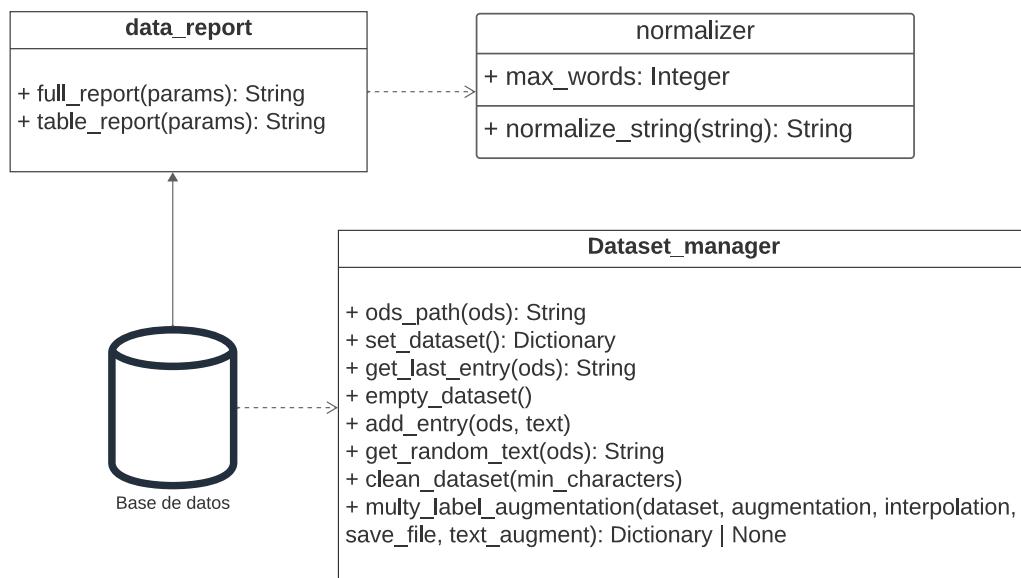


Fig. 3.2. Diagrama UML de la gestión de la base de datos

Módulo de gestión de la base de datos

Inicialmente se tenía la base de datos en forma de directorio, esto resultaba útil por razones explicadas posteriormente pero tenía limitaciones, una de ella era la gestión de la misma base de datos, siendo difícil inspeccionar todos los textos, eliminando así los duplicados y los vacíos. Esta es la funcionalidad que implementa este módulo, incorporando funciones para la gestión de la base de datos como lo son el borrado entero de la misma, la limpieza de textos vacíos y repetidos, el añadido de entradas, la extracción de una entrada aleatoria, la extracción de la última entrada o el aumento de la misma. Esta última funcionalidad será explicada en detalle mas adelante.

Módulo de tratamiento de texto

En este módulo se han implementado dos únicas funcionalidades pero ambas de alta importancia, una de ellas tienen como objetivo normalizar le texto a un número límite de caracteres por línea, haciendo que este sea legible en un archivo normal, esto resulta útil a la hora de generar un informe de una prueba en la que se incluyen los textos clasificados. La segunda de las funcionalidades es la de resumir los textos, esto también es útil a la hora de generar estas mismas salidas, sobre todo cuando el texto clasificado es muy largo como para poder leerlo rápidamente.

Módulo de generación de informe

Una vez se ha generado la base de datos este módulo itera por toda ella para extraer diversas métricas y generar un informe. Este puede ser de uno de dos tipos, textual y en forma de tabla, este primero contiene más información pero de una manera más extensa, mientras que la tabla contiene menos información pero esta es más fácil de digerir.

Ambos tipos de informe cuentan, por cada ODS el número de textos y la longitud máxima, mínima y media de estos, la única diferencia es la presentación de estos datos y que en el caso del informe textual se incluyen textos de ejemplo de cada objetivo.

3.1.3. Gestión de modelos

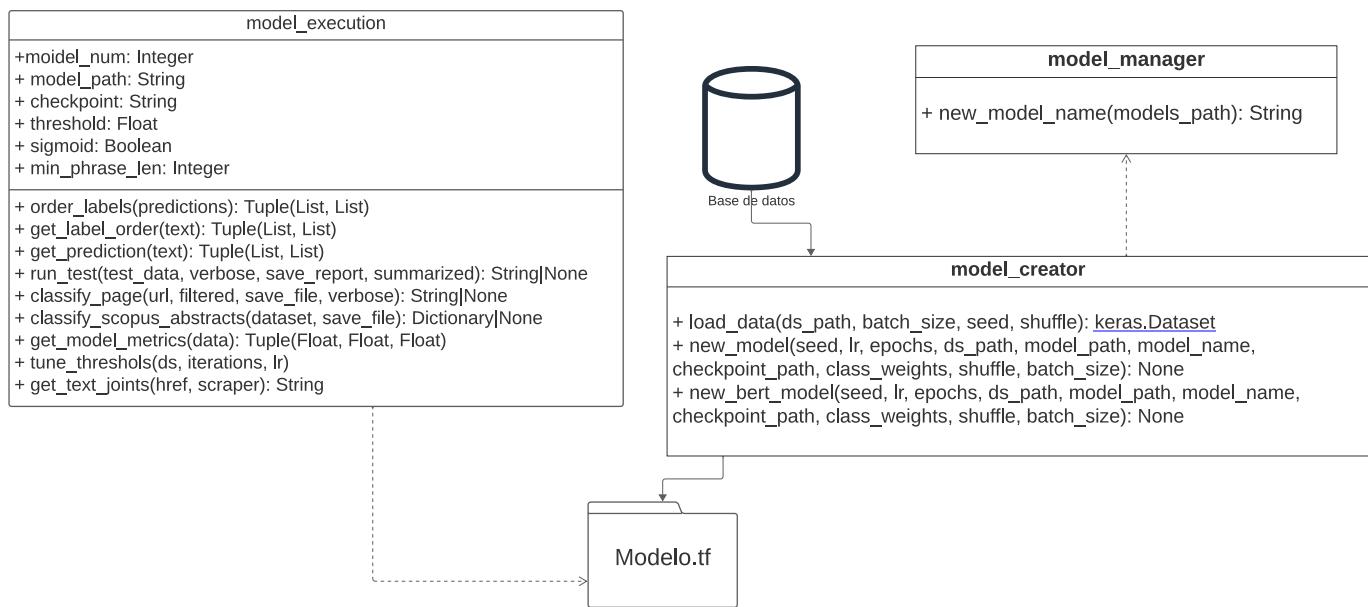


Fig. 3.3. Diagrama UML de la gestión de modelos

Módulo de creación de modelos

Este módulo es el responsable de crear, entrenar y validar los modelos. Soporta dos tipos de bases de datos como entrada, tanto en forma de directorio (explicado más adelante) como en un CSV y puede crear dos tipos de modelos uno basado en una red recurrente y otro basado en BERT, en ambos casos se introducen los mismos parámetros:

- Número de ciclos de entrenamiento: número de iteraciones de entrenamiento que realizar
- Tasa de aprendizaje: cuánta influencia tienen los errores en el entrenamiento, cuanto menor sea el número más ciclos de entrenamiento serán necesarios pero este ajustará mejor los pesos.
- Ubicación de la base de datos: directorio en el que encontrar los datos, tanto en forma de directorio como de archivo CSV
- Ubicación del modelo: directorio en el que guardar el modelo entrenado, por defecto se utiliza la regla de nombramiento “modelXX” donde “XX” es el número asociado y existe una función que determina dicho número automáticamente por lo que se suele dejar en blanco.
- Balanceo de clases: indica si se le deben asignar pesos a las clases en función del número de ocurrencias de cada una en la base de datos. Esto suele resultar en un modelo más uniforme pero puede interferir de manera significativa en el entrenamiento.

- Tipo de arquitectura: atributo utilizado para especificar si crear una red recurrente o un modelo basado en BERT.

El proceso de creación y entrenamiento es idéntico para ambos modelos, a excepción de la arquitectura usada, este comienza cargando la base de datos y dividiéndola en dos conjuntos, uno de entrenamiento y otro de validación, realizado una división del 70/20. Posteriormente se calcula el balanceo de clases, si este se ha especificado como activo, y finalmente se entrena el modelo. En última instancia se guarda el modelo y se utiliza el conjunto de test para calcular las métricas finales del modelo.

Módulo de ejecución

Una vez guardado un modelo no es del todo sencillo usarlo para clasificar textos, sobre todo por la necesidad de interpretar correctamente las salidas generadas. Este módulo codifica estas funcionalidades para poder usar los modelos fácilmente. La arquitectura de este módulo consiste en una única clase que representa un modelo, al instanciar dicha clase se carga el modelo especificado y se codifican datos necesarios para la ejecución como el orden de las clases en la salida de los modelos, el tipo de función de activación usada, el umbral a usar (explicado más adelante) y longitud mínima de frase.

Mencionar que las salidas generadas por los modelos tienen una forma de vector unidimensional con 17 valores entre 0 y 1, uno por cada objetivo. Dicho valor representa la probabilidad de que ese objetivo esté relacionado con el texto y puede ser interpretado de diversas maneras dependiendo de la salida usada por el modelo. La manera en la que se sabe a qué objetivo pertenece cada probabilidad es gracias a que la salida está ordenada siguiendo el orden alfabético de las clases, por lo que los porcentajes del vector de salida están ordenados de la siguiente manera [ODS1, ODS10, ODS11, ODS12, ..., ODS17, ODS2, ODS3, ..., ODS9]

La clase *model* cuenta con 10 métodos destinados a clasificar textos usando el modelo indicado e interpretar las salidas generadas, estos son:

get_label_order Este primer método es el central de toda la arquitectura, este coge un texto (o textos) dado, lo pasa por el modelo asignado y ordena la salida generada de mayor a menor probabilidad, devolviendo este vector de valores junto con uno ordenado de igual manera pero sustituyendo los valores por sus respectivos objetivos.

get_prediction Este segundo método encapsula al anterior, haciendo uso de `get_label_order` para generar los vectores de probabilidades de un texto dado y posteriormente interpreta estas salidas para generar la predicción final del modelo. Dicha interpretación depende de la función de activación presente en el modelo ya que esta cambia la forma del vector de probabilidades: si esta es sigmoid al texto se le asignan todos aquellos objetivos cuyos

porcentajes superen el umbral definido al instanciar la clase, en cambio, si la función de activación de salida es softmax este proceso es menos trivial y será explicado en un apartado posterior. Finalmente se devuelven los objetivos clasificados por el modelo con sus respectivas probabilidades.

order_labels Éste método no es crucial pero implementa una funcionalidad útil que se consideró que procedía encapsular en su función. Esta consiste en, dado un vector de probabilidades, devolver este ordenado de forma descendente junto con otro vector conteniendo los objetivos ordenados acorde a sus probabilidades.

run_test Método con una funcionalidad sencilla pero útil, dado una base de datos etiquetada clasificará todos los textos en ella y los comparará, pudiendo generar un informe en el que se incluyen todos los textos, los objetivos clasificados y sus porcentajes y finalmente la métrica de precisión total del modelo sobre ese conjunto de datos.

classify_page Una de las funcionalidades más sencillas pero curiosas de las implementadas es el hecho de clasificar una página web únicamente proporcionando la url, esto se hace mediante la librería de *requests* y *BeautifulSoup4* pudiendo extraer todo el texto de una página desde su código fuente.

classify_scopus_abstracts Este método es una primera aproximación de la funcionalidad final del modelo, clasificar abstracts extraídos de bases de datos académicas como Scopus o WoS. Este extrae los abstracts de un CSV, los clasifica todos y genera como salida el número de textos asignados a cada objetivo.

tune_threshold Más adelante se explicará en detalle el uso de umbrales, explicado superficialmente se usan para decidir qué etiquetas se asignan al texto y cuáles no en base a los porcentajes generados por el modelo. Este valor del umbral necesita ser ajustado y es por esto que por medio de un gran número de datos de prueba y calculando las métricas usando `get_model_metrics` para un alto número de valores del umbral se pueden identificar tendencias y relaciones entre las métricas y el valor del umbral, pudiendo así, decidir el valor óptimo de este.

get_text_joints Como añadido se definió este método que extrae texto de artículos de una determinada página, *jointsdgfound* [8], esto se hizo debido a que se tenía un gran número de enlaces a artículos de esta página clasificados con sus respectivos objetivos por lo que para clasificarlos mejor se optó por escribir este método, de esta manera se podían usar estos textos como conjunto preliminar de prueba.

Módulo gestor de modelos

Éste módulo se creó con la expectativa de una necesidad de gestionar los modelos mayor de la que finalmente resultó ser necesaria, es por esto que cuenta con una única función, el cálculo del nombre del siguiente modelo a entrenar, anteriormente se ha explicado que el nombre de los modelos siguen el patrón “modelXX” donde “XX” es un número que los identifica, esta función lee los modelos existentes y calcula cual es el siguiente número disponible para asignar.

3.2. Desarrollo realizado

Anteriormente se han definido todos los modelos desarrollados pero no como estos fueron creados. En este apartado se describirá el proceso de desarrollo seguido, como todas las partes interactúan y su papel en el sistema.

Para empezar cualquier trabajo de aprendizaje automático es fundamental contar con una buena base de datos ya que es de esta de la que se va a extraer el conocimiento a usar en el futuro. Esto se suele solucionar fácilmente buscando una base de datos disponible en internet o, hasta en algunos casos, generándola automáticamente de manera local. En este caso, después de una extensa investigación no se encontró ninguna base de datos existente con textos etiquetados de acorde a los objetivos de desarrollo sostenible. Descartando esta opción, la última alternativa disponible es generar esta base de datos de manera manual o pseudo-manual, en este caso se optó por utilizar técnicas de *Web Scraping* para extraer textos de páginas que contenían información segmentada acorde a los ODS.

Como primer recurso se acudió a la propia página de las naciones unidas para guardar las definiciones y metas de cada objetivo, esto probó ser una tarea imposible ya que esta información no se almacena de manera textual en la propia página si no que se encuentra codificada dentro de una serie de imágenes. Si se hubiera investigado mas seguramente existiera tal información almacenada de manera textual y accesible. Tratándose de textos comunes se consideró que lo más sencillo era buscar una segunda fuente, aquí es donde se encontró la página de *jointsdgfound* [8], esta contiene las definiciones, metas e información adicional sobre cada objetivo de una manera ordenada en la propia página por lo que su extracción y etiquetado fueron tarea sencilla. Se trajeron un total de 187 entradas, idealmente 11 por cada clase, pero los datos presentan una distribución altamente desbalanceada, siendo el que más textos asignados tiene el objetivo 17 con 20 datos y los que menos el 7 y 13 con 6.

Investigando se encontró un repositorio en esta misma página [8] conteniendo multitud de artículos, estando estos etiquetados con los objetivos sobre los que se habla. Dichos artículos se catalogaron, almacenando sus direcciones junto con los objetivos relacionados. Después de analizar estos datos nuevos se vio que la mayoría de ellos estaban relacionados con demasiados objetivos como para poder usarse para identificar las diferencias

entre ellos. Debido a esto no se usaron para entrenar ningún modelo pero si se almacenó la información recolectada para posteriormente usarse como una prueba más que realizar con los modelos.

Una vez recolectados los datos, se diseño un modelo de una red recurrente. Se eligió este tipo de arquitectura debido a su potencia a la hora de tratar con datos secuenciales, como lo son los textos. La arquitectura final del modelo implementado se basó en uno publicado por los desarrolladores de tensorflow, contando este con una serie de capas recurrentes conectadas a una red prealimentada. El resto de parámetros se eligieron en función del problema a resolver, la función de activación usada fue softmax, y como optimizador se eligió Adaptive Moment Estimation (ADAM), debido principalmente a su capacidad de adaptar la tasa de aprendizaje dependiendo del tipo de aprendizaje que se esté haciendo, y a que es ampliamente utilizado en tareas de aprendizaje automático debido a su versatilidad y potencia.

Una vez diseñado un modelo, este se entrenó usando los datos extraídos inicialmente. Este modelo primordial probó ser insatisfactorio (las métricas obtenidas se presentarán más adelante). Se conseguían unas métricas excelentes en el conjunto de entrenamiento de una manera muy rápida, mientras que las métricas del conjunto de validación se mantenían estables en un rango muy por debajo del aceptable. Este es un problema clásico de sobreaprendizaje, causado habitualmente por una falta de datos de entrenamiento.

Debido a estos resultados se buscó encontrar nuevas entradas, estas al final se extrajeron de los informes publicados anualmente por la Organización de las Naciones Unidas. Se tomó esta decisión ya que estos informes están segmentados por año y por objetivo, haciendo fácil su extracción cambiando la fecha en la dirección web, y su etiquetado, estando estos ya clasificados. La base de datos se aumentó hasta un total de 1118 entradas ya que se decidió dividir los informes por cada punto y añadir cada segmento resultante de manera independiente a la base de datos. De este modo se consiguió un conjunto de datos mucho más extenso y completo, y más importante aún bastante balanceado, con la única excepción del objetivo 17 ya que este cuenta con 120 entradas mientras el resto cuenta con alrededor de 60 cada uno.

Con estos nuevos textos añadidos se consiguieron los primeros resultados aceptables, siendo este un modelo bastante capaz, consiguiendo clasificar correctamente la mayoría de los casos y hasta razonando objetivos que no estaban asignados a los textos y que si tenían sentido clasificarse de esa manera. Este modelo, aunque no especialmente potente, sirvió como precedente, demostrando que si era posible diferenciar correctamente los objetivos y que los datos usados eran de cierta calidad.

De todas formas no era un modelo suficientemente capaz como para usarlo para realizar un estudio cuantitativo de una manera fiable, es por esto por lo que se decidió variar la arquitectura usada y entrenar modelos basados en BERT. Para esto se hizo uso de la librería tensorflow-hub la cual permite a los desarrolladores descargarse modelos preentrenados y adaptarlos a sus necesidades. Ese modelo descargado está formado por una

parte principal, esta, llamada *codificador*, emplea mecanismos de atención para codificar las entradas textuales, es en esta parte donde se asignan las importancias a las palabras y sus relaciones. Es por esto por lo que es necesario añadir una capa adicional que sea capaz de decodificar estas codificaciones para generar las salidas esperadas.

Aquí es donde se empezó a conseguir modelos realmente potentes, estos consiguen clasificar correctamente casi todos los textos y razonar objetivos nuevos con mucho más criterio. Esta última faceta es la más llamativa ya que inspeccionando las salidas se veía como el modelo era capaz de entender los textos de una manera profunda, interpretando patrones aún más complejos y clasificando objetivos nuevos a datos ya etiquetados con mucho más criterio.

Fue llegado este punto en el que se decidió implementar una clasificación multietiqueta, hasta hora el conjunto de datos de entrenamiento y de pruebas constaban de textos que tenían asignado únicamente un objetivo, *one-hot* y en la interpretación de las salidas solo se tenía en cuenta el objetivo con mayor porcentaje asignado.

Aquí se enfrentó uno de los principales problemas encontrados en el desarrollo, la función de activación de salida usada hasta ahora era softmax, esta era la decisión correcta debido al formato de los datos, pero resultaba ser un problema a la hora de implementar esta nueva funcionalidad.

Como solución se decidió implementar una nueva forma de interpretar las salidas. Esta es similar a la que se utiliza con redes con función de activación sigmoid, en este caso se asignan todas las clases cuyo porcentaje supere un umbral. La técnica implementada es similar ya que asignan todas las clases que superen un umbral, la diferencia clave está en la definición de dicho umbral.

En el caso de las redes con función de activación sigmoid simplemente se define un porcentaje mínimo a usar como umbral. En el caso de la función softmax esta solución no resultaría muy efectiva ya que la magnitud de los porcentajes generados varía ampliamente debido a su naturaleza aditiva. Esto hace que en los casos en los que haya muchos objetivos relacionados con un texto, la probabilidad estará distribuida entre muchas clases, siendo esta relativamente baja en todas pero, en un caso ideal, relativamente similar en magnitud entre todas las clases relacionadas. El caso opuesto no resultaría en un problema ya que si solo hay un objetivo relacionado el porcentaje de este será alto mientras que la probabilidad restante estará dividida entre el resto de clases.

Son estas características las que se aprovechan en esta nueva técnica, en esta se define un umbral inicial, este, una vez se obtienen las probabilidades generadas por el modelo se multiplica por la mayor de todas, de esta manera se genera un umbral adaptado al dato de entrada. Si este está relacionado con diversos objetivos y se genera un vector con probabilidades similares entre varias clases, estas serán similares en magnitud pero pequeñas, generando, a su vez un umbral final bajo, consiguiendo así clasificarlas. Por otra parte si el texto solo está relacionado con un objetivo, o es uno muy dominante, este obtendrá un porcentaje asignado alto por lo que no se asignarán más objetivos.

Esta funcionalidad resultó funcionar según lo esperado, clasificaba correctamente más de un objetivo en algunos casos, pero se vio que era imprescindible realizar un buen ajuste del umbral (De aquí en adelante se utilizará umbral para referirse al porcentaje por el cual se multiplica la probabilidad mayor para obtener el umbral de verdad) ya que es necesario encontrar un buen valor para realizar una buena clasificación.

Es aquí donde se llegó a uno de los puntos de inflexión, se consideró terminar el desarrollo, con un buen modelo clasificador y adaptado para realizar una clasificación multi-etiqueta pero el método usado, al ser una variación de otro, no resultaba cómodo de usar como modelo final. Después de investigar se decidió realizar la clasificación multi-etiqueta de una manera mas aceptada y probada. Haciendo uso de una función de activación de salida sigmoid. Esta función, como se explicó en el estado de la cuestión, genera un vector de probabilidades independientes por lo que se pueden asignar todas aquellas que superen un porcentaje determinado, denominado umbral.

Se entrenó un nuevo modelo con los datos ya existentes, pero al no tener estos una forma adecuada no se obtuvieron resultados diferentes a los que se consiguieron con el modelo anterior. Esto es debido a que todos los datos tienen un único objetivo asignado, no hay ningún texto con más de un objetivo asociado por lo que, independientemente del tipo de arquitectura y función de activación usada, el modelo no va a ser capaz de clasificar más de un objetivo porque no es un caso al que haya sido expuesto previamente.

Esto es algo que se pensó solucionar a posteriori, a la hora de clasificar un texto más grande dividir este en sus diferentes frases, para posteriormente usar un modelo especializado en identificar objetivos uno a la vez y en frases cortas para clasificar cada una de manera independiente y finalmente realizar la media de los porcentajes generados para cada una. De esta manera si se habla de más de un objetivo esto se verá reflejado en alguna frase, teniendo un impacto en los porcentajes finales y, si todo funciona según lo esperado, clasificando el texto correctamente. Esto probó no merecer la pena, por un lado, dividir un texto en sus frases y clasificarlas independientemente resulta tener un impacto significativo en el rendimiento, tardando bastante más que clasificando textos enteros. Por otro lado analizando los textos con los que se va a usar el modelo final, textos científicos cortos (abstracts), se vió que resultaría más útil generar un modelo capaz de clasificar correctamente textos enteros que las diferentes frases, ya que estos textos suelen tener menos información y esta está más repartida. Adicionalmente, de esta manera también se aprovecha la naturaleza bidireccional de los modelos basados en BERT. Es por esto que se decidió generar una base de datos con múltiples etiquetas.

Para conseguir generar una base de datos multi-etiqueta se hizo uso de una técnica muy común en el aprendizaje automático, el aumento de datos, aplicándola de una manera particular. Esta técnica consiste en generar datos nuevos a partir de unos ya existentes, variando sutilmente estos mismos de tal manera que son datos diferentes pero no lo suficiente como para que su etiqueta asociada ya no tenga sentido. Esta técnica es usada en multitud de problemas, el más llamativo de todos siendo en el uso de imágenes, pudiendo

rotar, desplazar, recortar o hasta alterar los colores de la imagen sin que esta pierda su sentido. Cuando se trata con datos textuales existen multitud de formas de aumentar estos datos, una serie de ellos se demonima Easy Data Augmentation (EDA), aumento de datos sencillo. Esta serie de algoritmos realiza operaciones como sustituir palabras por sinónimos, eliminar palabras aleatorias, cambiar de posición una serie de palabras de manera aleatoria. Adicionalmente existen una serie de métodos algo más sofisticados como son:

- Wordnet: Este es el nombre de uno de los modelos de vectorización de texto mas potentes, este se usa para aumentar la base de datos de manera similar a cuando se reemplazan palabras con sus sinónimos pero en vez de elegir estos sinónimos en base a una base de datos estas palabras sinónimas son aquellas que se encuentran próximas a la palabra objetivo en este espacio vectorial, de esta manera, a ojos del modelo las palabras no serán las mismas pero su significado si será parecido.
- Traductor: Este tipo de aumento es muy curioso pero útil a la vez. Este se basa en algo que todo el mundo ha hecho, escribir una frase en un idioma, traducirla a otro cualquiera y volver a traducir esa frase al idioma original. El resultado rara vez es la misma frase pero el significado sigue siendo el mismo en la mayoría de los casos. Este tipo de aumento hace uso de este proceso para traducir la frase a un idioma determinado y luego de vuelta al original, de esta manera obtendremos dos frases diferentes pertenecientes a la misma etiqueta.

Esta técnica de aumento se usó para generar un nuevo conjunto de datos multi-etiquetados partiendo de los anteriores. Como técnica de aumento a usar se escogió Wordnet. Este aumento consistió en elegir de entre 2 y 5 textos, cada uno de un objetivo diferente, estos textos se transformaron usando Wordnet y se juntaron en una entrada nueva, estando esta etiquetada con los objetivos de los textos seleccionados. De esta manera se aumentó la base de datos a un total de más de 5000 entradas.

Con este nuevo conjunto de datos se entrenó un nuevo modelo, el cual rindió mejor que los anteriores en tareas de clasificación multi-etiqueta.

Finalmente, investigando trabajos similares se descubrió una página que contaba con una gran cantidad de artículos relacionados con los objetivos de desarrollo sostenible, en la mayoría de los casos con mas de uno a la vez, el *International Institute for Sustainable Development* (IISD) [44]. Después de examinar estas publicaciones se vio que eran de una buena calidad, y contaban con el aval de haber sido usadas en trabajos similares. Fue por estas razones por la que se decidió generar un último conjunto de datos, no tan grande como el generado con técnicas de aumento de datos, pero de mayor calidad. Se desarrolló un último script de extracción de datos que recogía los enlaces y otro que extraía los textos de dichos enlaces y los guardaba con sus respectivas etiquetas. Fue en este momento cuando se llevó a cabo una transición de una base de datos en forma de directorio a una en un archivo CSV, ya que la naturaleza multi-etiqueta hace que esta última sea la mejor opción.

Como último añadido se vio que la página del United Nations Development Programme (UNDP) [45] contenía definiciones diferentes y extensas de los objetivos por lo que también se añadieron a la base de datos

Se desarrollaron dos últimos grupos de modelos, unos entrenados únicamente con los datos extraídos de las publicaciones de IISD y otros finales desarrollados con el conjunto final, contando con los datos extraídos al inicio (las definiciones y los informes publicados por naciones unidas) y estas últimas dos extracciones. De este último conjunto se dejaron fuera los datos creados mediante técnicas de aumento ya que se consideró que el tamaño de este era suficiente y se priorizó la calidad de los datos.

Estos últimos modelos son los que mejores resultados obtuvieron, no en magnitud de algunas métricas si no en el equilibrio de estas, obteniendo buenos resultados en todas.

Finalmente se utilizó el modelo para realizar una investigación del contenido de las publicaciones científicas publicadas, esta investigación se cubrirá en detalle en apartados posteriores.

4. VALIDACIÓN Y PRUEBAS

Como parte esencial de cualquier desarrollo de un sistema de aprendizaje automático está el apartado de validación y pruebas, es aquí donde se ve si realmente el trabajo llevado a cabo ha sido fructífero. Y más importante aún, se aprecia que cambios y desarrollos son los que mayor impacto han tenido.

Al final se crearon un total de 23 modelos, 10 de ellos basados en una red neuronas recurrente mientras que el resto están basados en BERT. De los 23 modelos finales, 5 de ellos cuentan con una función de activación de salida softmax mientras que el resto cuenta con una sigmoid. Los modelos restantes se dividen en tres grupos dependiendo del conjunto de datos con los que fueron entrenados, están los que se entrenaron con los datos resultados del aumento multi-etiqueta, aquellos entrenados únicamente con los textos extraídos de IISD [44], la última página web mencionada, y por último aquellos que fueron entrenados juntando todos estos conjuntos, a excepción del aumentado que se descartó por motivos de consistencia.

En total se han llevado a cabo tres tipos de pruebas, la mas extensa y fiable de ellas se ha realizado sobre el conjunto de prueba seleccionado. Otro de ellos se realizó clasificando textos extraídos manualmente y viendo si los resultados cuadran con textos correctamente y finalmente se han clasificado multitud de abstracts de publicaciones científicas alojados en la página WoS, estos fueron extraídos siguiendo una serie de consultas publicadas en la literatura, las cuales afirmaban seleccionar trabajos de únicamente un ODS cada una.

4.1. Conjunto de pruebas

Las métricas expuestas en este apartado han sido calculadas clasificando los textos del conjunto de pruebas final, la mayoría de los cuales no han sido vistos por ningún modelo anteriormente. Esto se debe a la naturaleza secuencial del desarrollo, en la última fase se juntaron todos los textos extraídos y se realizó una división 75/25 entre el conjunto de entrenamiento y el de pruebas. Es probable que alguno de los textos usados para el entrenamiento en modelos anteriores acabaran en el conjunto de pruebas final, pudiendo alterar los datos recogidos del mismo. La probabilidad de que esto ocurra con una magnitud tal como para alterar estas métricas de una manera significativa es relativamente baja por lo que estas se asumen como verdaderas y fiables.

En este apartado se muestran las siguientes métricas: precisión, exhaustividad, exactitud y valor-F1, estas se calcularon en base a la cantidad de verdaderos positivos (tp), verdaderos negativos (tn), falsos positivos (fp) y falsos negativos (fn) predecidos por el modelo de la siguiente manera:

$$precision = \frac{tp}{tp+fp}$$

$$\text{exhaustibilidad} = \frac{tp}{tp+fn}$$

$$\text{exactitud} = \frac{tp+tn}{tp+fp+tn+fn}$$

$$\text{valor - F1} = 2 * \frac{\text{precision} * \text{exhaustibilidad}}{\text{precision} + \text{exhaustibilidad}}$$

El conjunto de estas métricas puede revelar la correcta funcionalidad de un modelo en tareas de clasificación de una manera profunda y completa ya que se cubre todo el rango de posibilidades.

Con la *precisión* se calcula la capacidad que tiene el modelo para predecir clases correctas, esto es, que fracción de las clases que se asignan como positivas lo son de verdad.

Con la *exhaustividad* se comprueba lo contrario que con la precisión, esto es cuántos de los datos que son positivos son clasificados como tal por el modelo. De esta manera se puede saber que fracción de los datos no es clasificada por el modelo.

Con el *valor-F1* se combinan las métricas de precisión y exhaustividad en una sola. Tiene especial importancia si ambas métricas son igual de relevantes, afirmación que no siempre es verdadera, dependerá exclusivamente del objetivo y funcionalidad esperada del modelo.

Finalmente la *exactitud* mide la fracción de clases que han sido asignadas correctamente, tanto las positivas como las negativas. La fiabilidad de esta métrica depende de varios factores, uno de los cuales es en nivel de balanceo de las clases, si estas están desbalanceadas esta métrica se podrá inflar clasificando la clase mayoritaria como verdadera siempre.

En el caso de este proyecto, pese a que se han calculado todas las métricas, las más importantes son la exactitud y el valor-F1. Se han seleccionado estas dos ya que en conjunto representan todos los aspectos relevantes del modelo. La exactitud es muy útil y es ampliamente usada como métrica pero hay que usarla con cuidado y en caso de duda contrastarla con el valor-F1. Esto es debido a que, como se ha mencionado, depende del nivel de balanceo de los datos usados. En el caso de los conjuntos de entrenamiento y pruebas, y como se puede ver en la TABLA 4.2, se obtiene, en conjunto, una desviación estándar de 466. Siendo esta relativamente alta para el número de datos total usados, representados en la TABLA 4.1. Como se aprecia en dicha tabla, las clases más desbalanceadas son la del objetivo 17, contando con 2521 datos totales, y la del objetivo 4, contando con 552, Más del doble de la media de 1070.65 en el primer caso y aproximadamente la mitad en el segundo. Si se obvian estas dos clases, se obtiene una desviación media de 291.53, un valor mucho mas razonable.

Es debido a ese desbalanceo por lo que se ha seleccionado el valor-F1 como una métrica relevante. De esta manera se puede validar la fiabilidad de la exactitud obtenida por un modelo analizando si esta es fruto de una clasificación defectuosa y sesgada y realizar, de esta manera, un análisis más profundo de los mismos.

TABLA 4.1. NÚMERO DE DATOS POR OBJETIVO

Modelo	Ocurrencias
ODS1	791
ODS2	925
ODS3	1121
ODS4	552
ODS5	787
ODS6	775
ODS7	875
ODS8	1076
ODS9	991
ODS10	925
ODS11	750
ODS12	1127
ODS13	1969
ODS14	940
ODS15	1218
ODS16	858
ODS17	2521
TOTAL	18201

TABLA 4.2. MÉTRICAS ESTADÍSTICAS DE LA BASE DE DATOS FINAL

Métrica	Valor
Media	1070.65
Mediana	925.0
Desviación media	466.86

A continuación se muestra la evolución de la métricas de los diferentes modelos a medida que se va aumentando el valor del umbral, este valor, como se definió en la solución propuesta, representa la facilidad con la que el modelo clasificará más de una clase a un mismo texto, siendo un 0 el caso extremo en el que se clasificarán todas las clases y un 1 el caso en el que, para los modelos con función de activación softmax solo se asignará una etiqueta a un texto, y para los modelos con función de activación sigmoid no se clasificarán casi ninguna etiqueta ya que es muy raro que esta tenga un 100 % de posibilidad asignada.

A continuación se muestran las métricas obtenidas por los diferentes modelos sobre el conjunto de datos de prueba final. Destacar la importancia de analizar el eje de abscisas en cada gráfica, puesto que este empieza en valores diferentes dependiendo de la gráfica, por lo que dos gráficas similares pueden variar en magnitud.

4.1.1. Métricas redes recurrentes

En primer lugar se muestran los datos obtenidos por los modelos basados en redes recurrentes.

Los resultados de estas pruebas son bastante similares entre sí, variando únicamente y de manera leve en magnitud por lo que a continuación se incluyen las dos gráficas más relevantes, el resto se obvian por motivos de reiteración.

En estas dos gráficas, figuras 4.1 y 4.2, se puede apreciar una distribución ciertamente horizontal, con poca variación de las métricas en relación al valor del umbral. Adicionalmente se observa un valor medio del 70 % de exactitud, de 40-50 % de precisión y de menos de 30 % de exhaustividad y valor-F1. Esto es lo que se esperaría de estos modelos ya que están diseñados y entrenados con datos con una sola etiqueta, los denominados como *one-hot* y no se espera un buen rendimiento de ellos en tareas de clasificación multi etiqueta. Adicionalmente un valor tan bajo de precisión indica una capacidad pobre de los modelos para asignar todos los objetivos relacionados con un texto. De todas formas consiguen un valor aceptable de exactitud por lo que sirven como una buena base sobre la que desarrollar modelos más potentes y comparar los resultados.

Metricas modelo 1

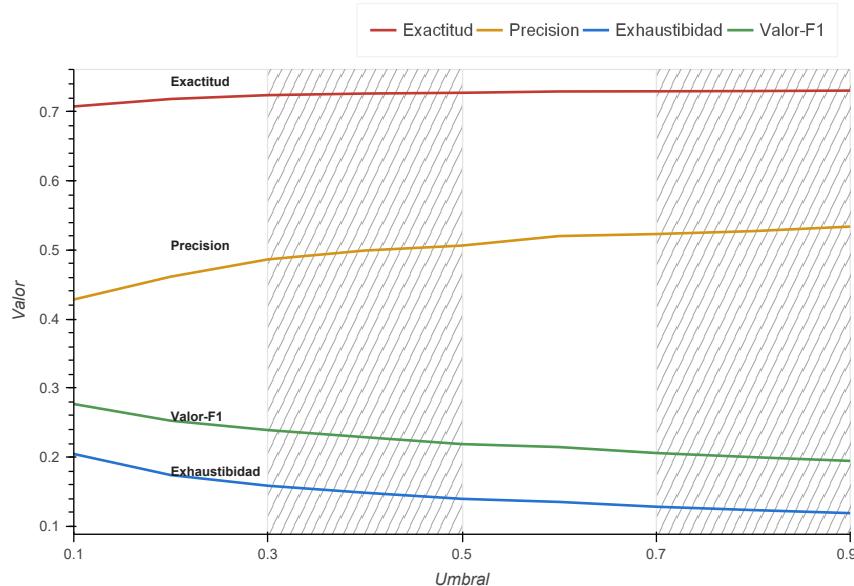


Fig. 4.1. Métricas modelo 1

Métricas modelo 2

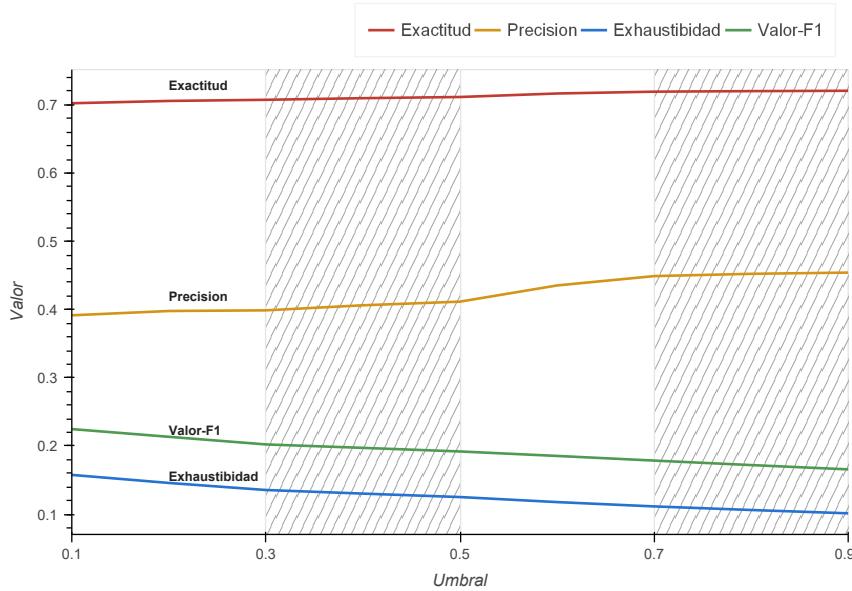


Fig. 4.2. Métricas modelo 2

4.1.2. Métricas modelos BERT

A continuación se muestran las métricas de aquellos modelos basados en BERT.

Estos modelos a su vez se han dividido en tres grupos dependiendo del conjunto de datos que se ha usado para entrenarlos.

Datos iniciales - *one-hot*

En primer lugar se encuentran aquellos entrenados con el primer conjunto de todos, aquel que cuenta únicamente con una etiqueta por dato.

De todos los modelos entrenados con estos datos, los primeros a analizar con aquellos con función activación softmax, estos son los modelos 11-13-14.

Como se puede apreciar en las gráficas, figuras 4.3 a 4.5, se ve como los modelos tienen un rendimiento estable, es decir, este no varía casi con el umbral. Esto es debido a la función de activación usada ya que esta hace que no suelan asignar muchas clases a las entradas, obteniendo así una exhaustividad relativamente baja y esta no varia al pasar a valores de umbral mayores que un 0.3 puesto que, pasado este punto, la diferencia entre los porcentajes precedidos por el modelo es tal que no se suelen asignar muchos. Esto nos indica que las salidas de estos modelos suelen ser correctas, pero también se suelen dejar objetivos asociados sin asignar. Esto entra en lo que cabía esperar de este tipo de modelos.

Métricas modelo 11

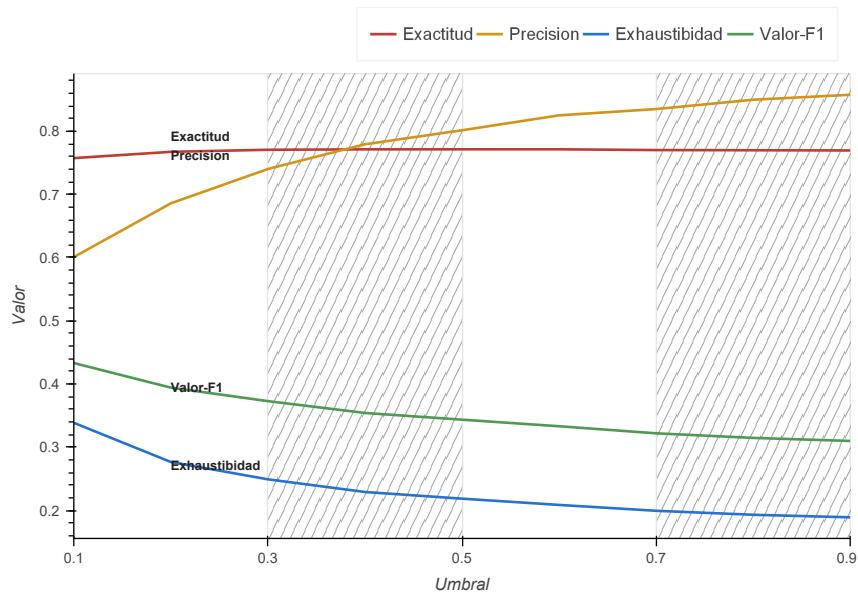


Fig. 4.3. Métricas modelo 11

Métricas modelo 13

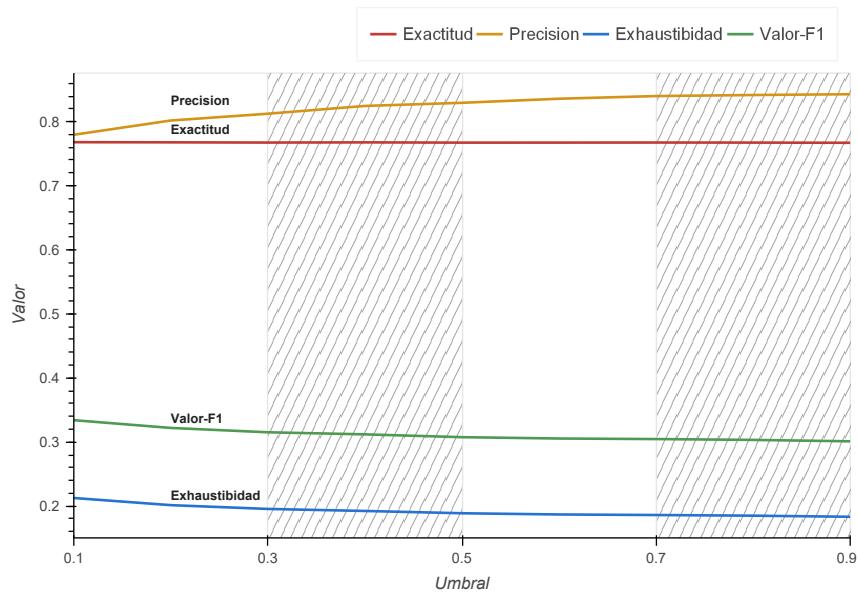


Fig. 4.4. Métricas modelo 13

Métricas modelo 14

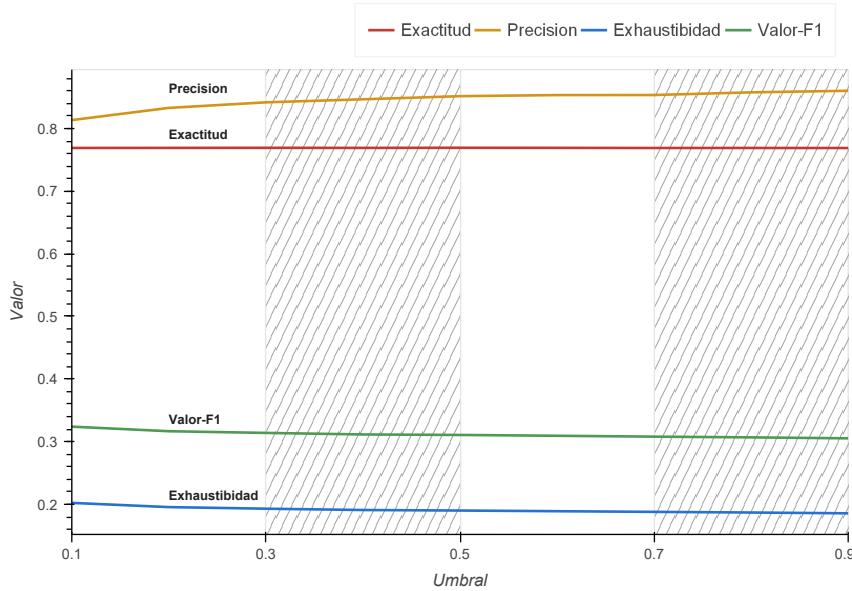


Fig. 4.5. Métricas modelo 14

El resto de modelos son aquellos que cuentan con una función de activación softmax.

En las gráficas, figuras 4.6 a 4.8, se puede apreciar que el rendimiento de estos modelos, en la mayoría de los casos, es similar a los anteriores, obteniendo unos buenos resultados en precisión y exactitud pero unos resultados mediocres en exhaustividad y valor-F1. Esto era de esperar debido a que están entrenados con el mismo conjunto de datos por lo que aprenderán de la misma manera. Destacar que el modelo 16, figura 4.6, tiene una gráfica con una forma diferente, esta presenta unos resultados más equilibrados, pero a costa de disminuir la magnitud de los mismos. Esto es debido a que fue diseñado con una función de loss⁴ diferente, siendo esta poco adecuada para entrenar modelos con función de activación softmax, en concreto se usó *Categorical cross entropy*, estando esta función diseñada para datos con una codificación *one-hot* como los usados con funciones de activación sigmoid.

⁴Una función de pérdida que mide la discrepancia entre las predicciones de un modelo y los valores reales durante el entrenamiento.

Métricas modelo 16

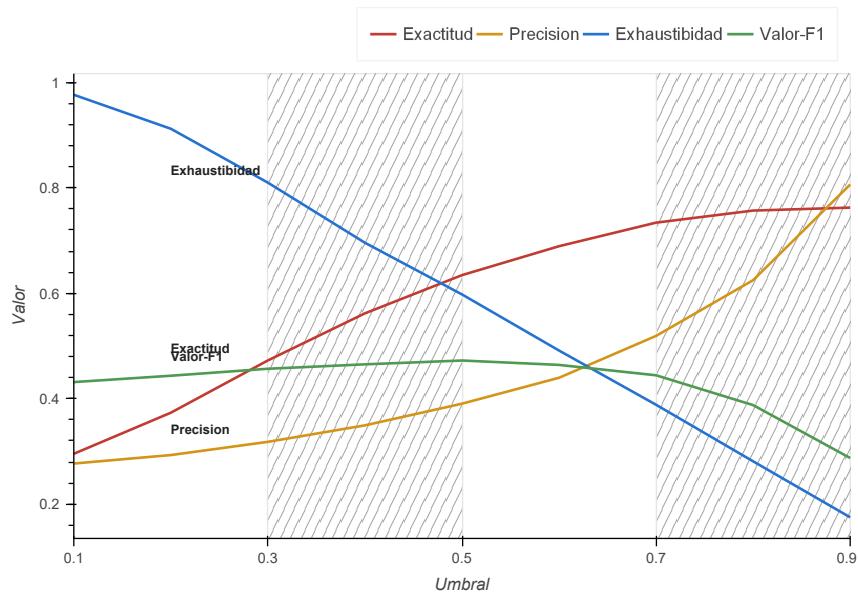


Fig. 4.6. Métricas modelo 16

Métricas modelo 17

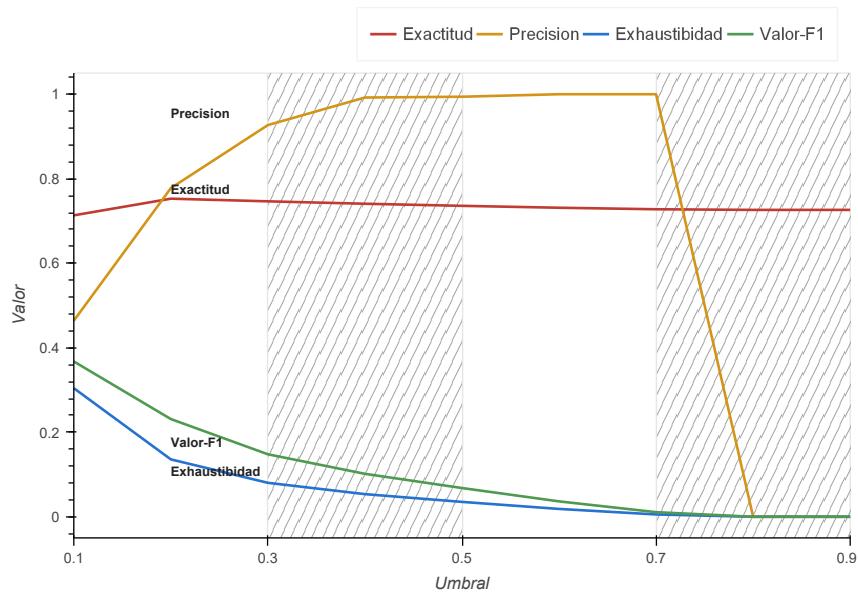


Fig. 4.7. Métricas modelo 17

Métricas modelo 18

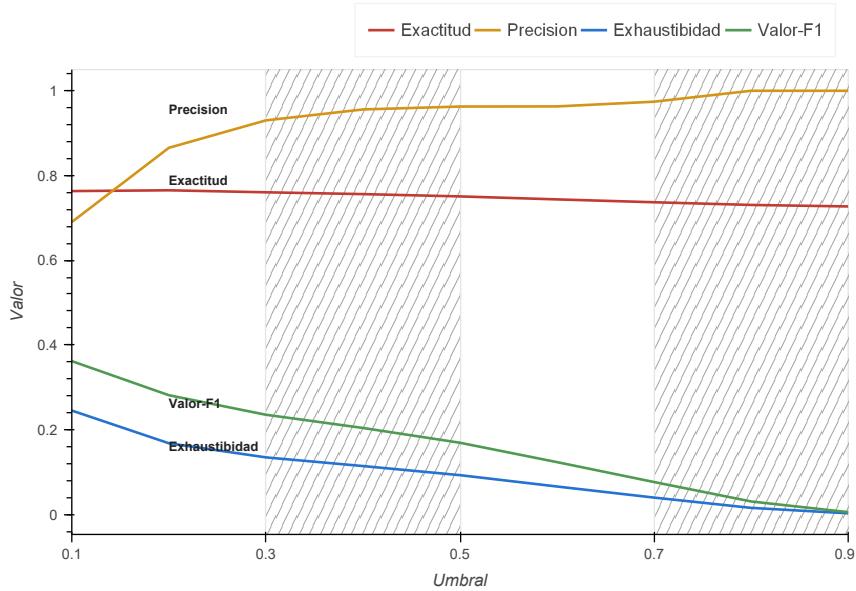


Fig. 4.8. Métricas modelo 18

Datos aumentados - multi-etiqueta

Los siguientes modelos fueron entrenados con el conjunto de datos aumentado:

En las gráficas, figuras 4.9 a 4.11, se puede apreciar que, pese a una gran cantidad de datos nuevos multi-etiquetados no obtiene unos resultados generales mucho superiores en cuanto a la magnitud de los mismos, de todas formas estos son más equilibrados, siendo superiores en las tareas de clasificación multi-etiqueta, como se observa en las métricas de exhaustividad y valor-F1, estando estas entre 10 y 20 puntos porcentuales por encima de los anteriores modelos. Adicionalmente con valores altos de umbral se consigue una precisión extremadamente alta, estando cerca del 100 %, indicando esto que las etiquetas a las que se les asigna una probabilidad muy alta son en su mayoría correctas.

Métricas modelo 19

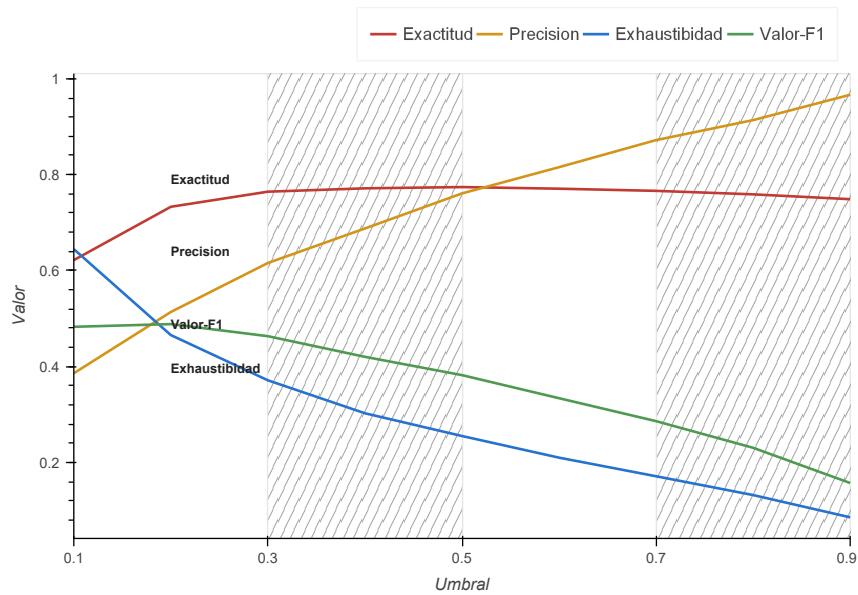


Fig. 4.9. Métricas modelo 19

Métricas modelo 20

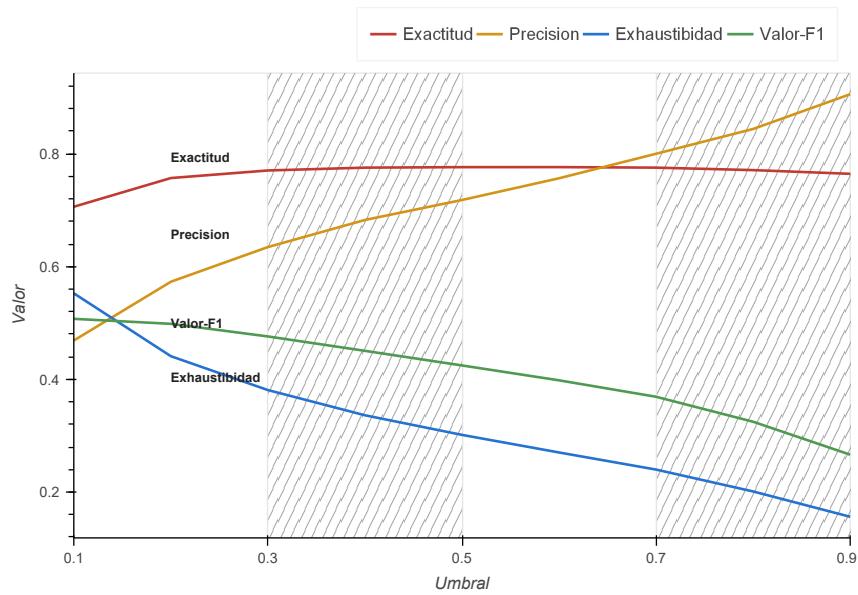


Fig. 4.10. Métricas modelo 20

Métricas modelo 21

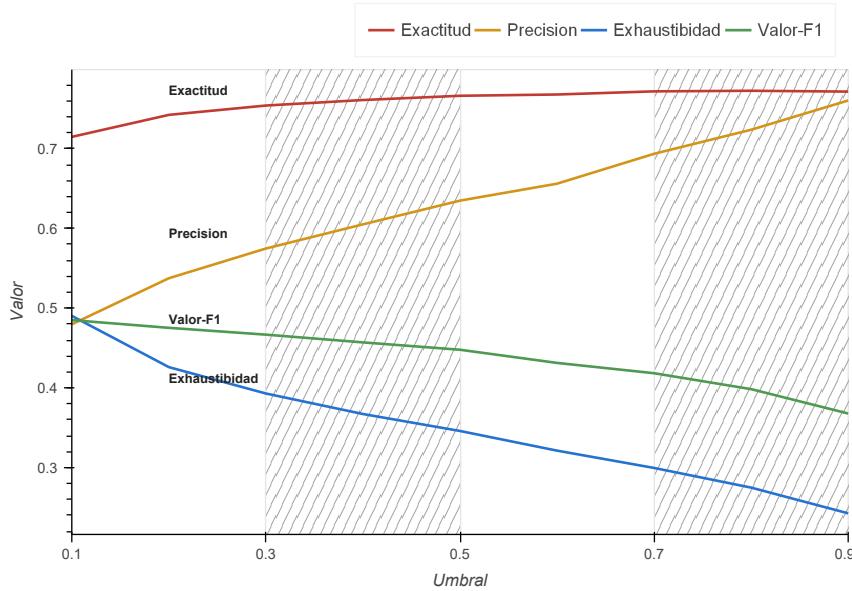


Fig. 4.11. Métricas modelo 21

Datos extraídos de publicaciones en iisd - multi-etiqueta

A continuación se creó una nueva base de datos extrayendo las publicaciones de IISD, esta tiene una naturaleza multi-etiqueta, se entrenó un único modelo que obtuvo unos resultados realmente buenos y equilibrados como se observa en las métricas de la gráfica, figura 4.12. Es esta se puede apreciar como por primera vez se obtiene unos valores buenos en las métricas de exhaustividad y valor-F1. Adicionalmente se observa que para un valor del umbral de al rededor de 0.25 se obtienen unas métricas realmente buenas en general. Esto puede parecer tentador pero se aprecia que el valor de la precisión es más bajo de lo habitual, por lo que estas métricas tan buenas de valor-F1 y de exhaustividad se obtienen a causa de una precisión más baja, por lo que se asignan más objetivos de los adecuados. Esto se considera menos ideal que el caso contrario, en el que se asignan menos etiquetas pero con la garantía de que estas si están relacionadas.

Destacar que este modelo ha sido entrenado con algunos de los datos usados para las pruebas por lo que las métricas pueden estar ciertamente infladas pero independientemente de esto las métricas obtenidas son positivas.

Métricas modelo 22

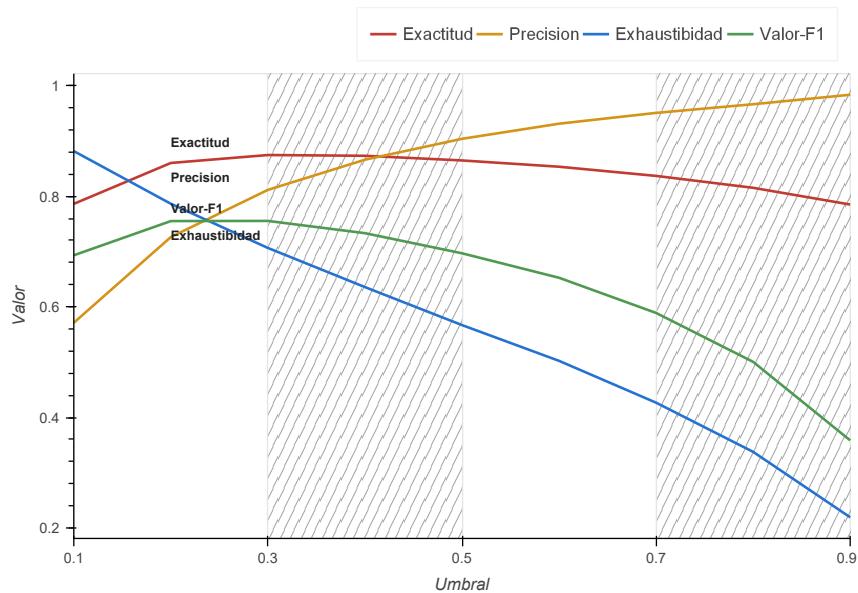


Fig. 4.12. Métricas modelo 21

Datos finales - multi-etiqueta

Finalmente se generaron una serie de modelos con la base de datos conjunta, estos modelos son los únicos que no se han entrenado con ninguno de los datos incluidos en el conjunto de pruebas usado para el cálculo de las métricas, por lo que estas son las más fiables de todos los modelos.

Observando las gráficas, figuras 4.13 a 4.16, se aprecia que estas presentan distribuciones similares, contando con unos resultados medios en todas las métricas para umbrales bajos y estas se van dividiendo a medida que este aumenta, siendo el punto medio óptimo aquel asociado a un valor del umbral de 0.5, apreciable en mayor medida en la figura 4.16, donde todas las métricas menos la de la exactitud se cruzan en este valor, para el resto de modelos esta afirmación es, en cierta medida, arbitraria pero indudablemente es un punto óptimo para la mayoría.

Metricas modelo 23

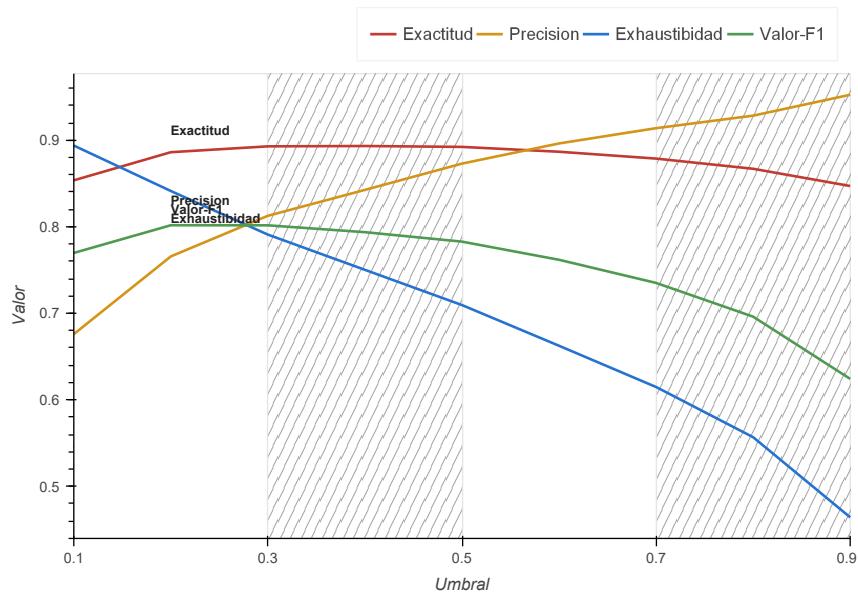


Fig. 4.13. Métricas modelo 23

Metricas modelo 24

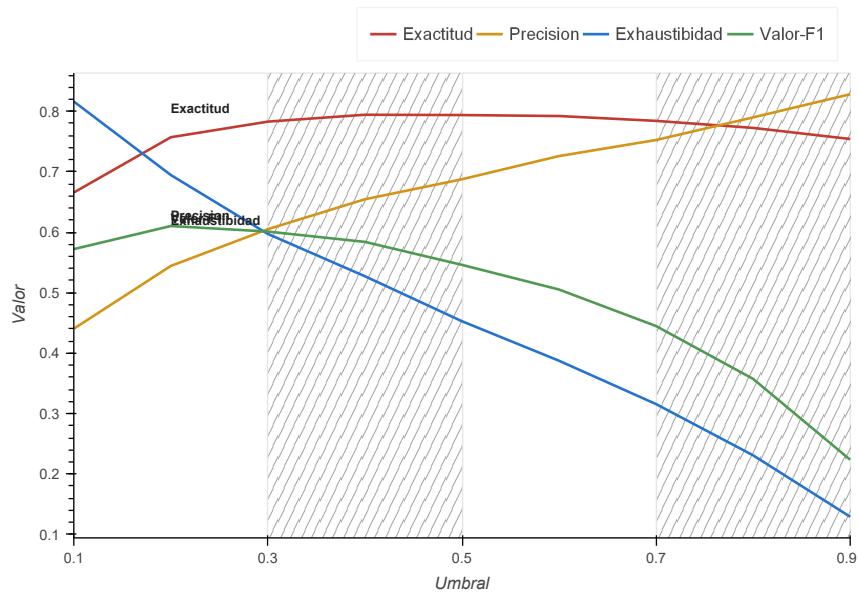


Fig. 4.14. Métricas modelo 24

Métricas modelo 25

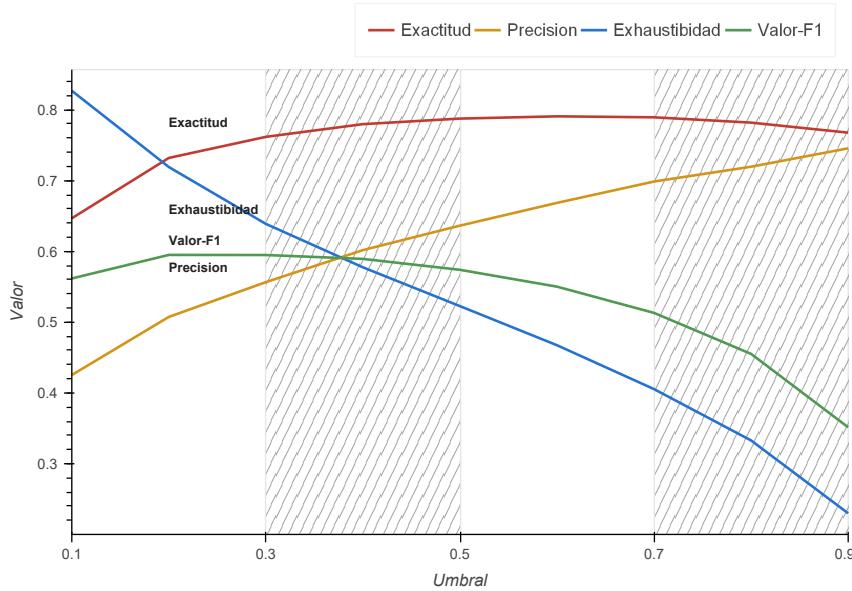


Fig. 4.15. Métricas modelo 25

Métricas modelo 26

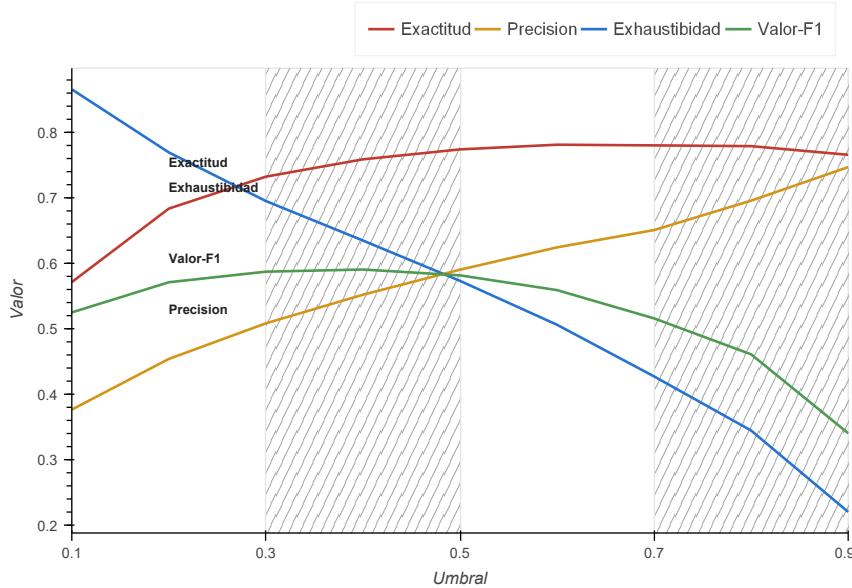


Fig. 4.16. Métricas modelo 26

Como conclusión, en estas gráficas, figuras 4.3 a 4.16, se puede apreciar que para un valor medio del umbral se encuentra el punto de mayor equilibrio entre las métricas. Es habitual encontrar afirmaciones indicando que un valor de 0.5 es lo adecuado cuando se trabaja con salidas sigmoid y una clasificación multi-etiqueta. Esto se ve reforzado visualizando la figura 4.16 ya que tres de las 4 métricas se cruzan exactamente en este valor. Para el resto no resulta tan fácil apoyar esta afirmación pero si se priorizan los valores de precisión y en menor medida, el valor-F1, parece que otra vez este valor de

umbral es el que consigue unos resultados equilibrados.

Cuando se habla de las funciones de activación softmax este valor se ve menos claro, observando las gráficas podría parecer intuitivo darle un valor relativamente bajo pero esto no parece una decisión adecuada ya que se estaría siendo muy poco restrictivo sin ningún resultado positivo aparente reflejado en las métricas. Debido a esto y por razones de simplicidad se usará un umbral de 0.5 para todos los modelos, apoyado en el hecho de que las gráficas, figuras 4.3 a 4.5, no presentan apenas variación en las métricas en relación al umbral.

4.1.3. Métricas finales

Finalmente, seleccionando un umbral medio de 0.5 y calculando la exactitud y valor-F1 de todos los modelos sobre el conjunto definido de pruebas se obtienen las siguientes gráficas de exactitud, figura 4.17, y valor-F1, figura 4.18, comparando el rendimiento de todos los modelos.

Destacar que esta primera gráfica, figura 4.17, tiene el eje de ordenadas, representando el valor de exactitud, truncado, iniciando este en 0.5. Esto es debido a que las diferencias entre los modelos no son tan grandes como para ser apreciables en una escala más grande.

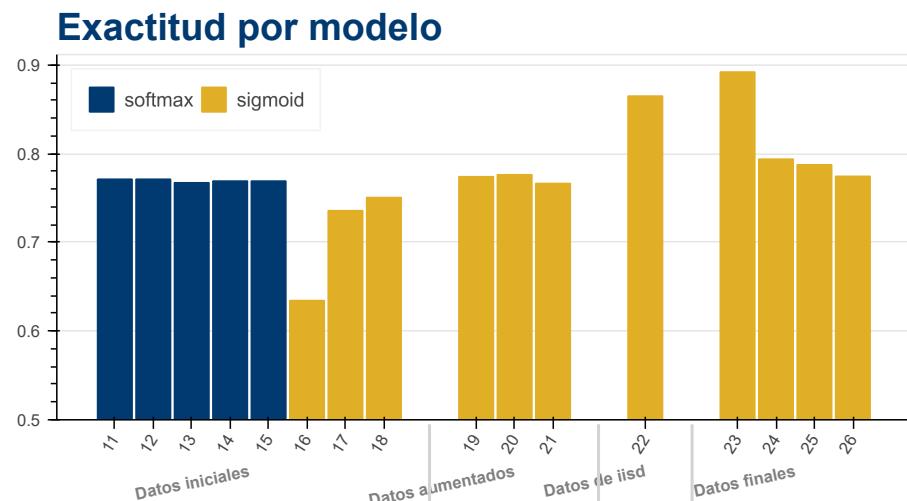


Fig. 4.17. Exactitud de todos los modelos

Valor-F1 por modelo

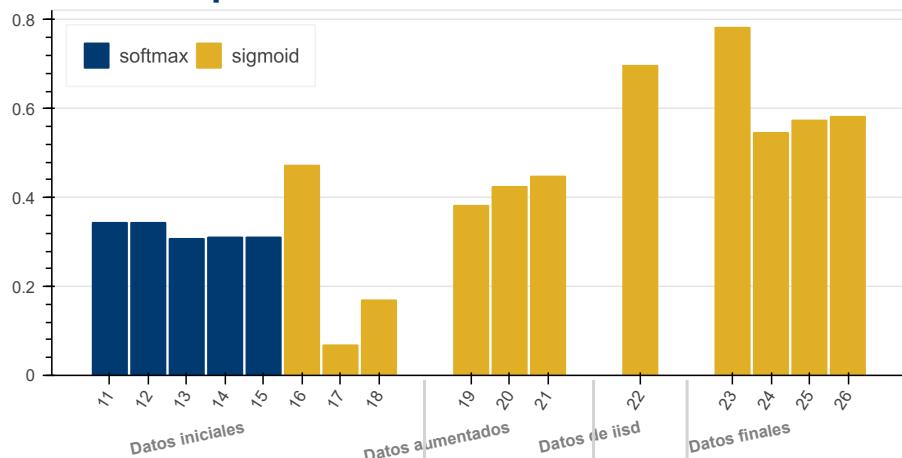


Fig. 4.18. Valor-F1 de todos los modelos

En la gráfica, figura 4.17, se puede observar que la exactitud obtenida por los diferentes modelos es relativamente parecida, rondando en torno al 80 % en la mayoría de los casos, con alguna que otra excepción. Tendencia que se mantiene al analizar el valor-F1, en la figura 4.18 aunque en menor magnitud, rondando en este caso un valor del 50 %.

También se puede observar que los modelos con función de activación sigmoid pero entrenados con el primer conjunto de datos obtienen un rendimiento peor que los modelos entrenados con los mismos datos pero con función de activación softmax. Esto puede ser resultado del tipo de datos usados ya que la función softmax está diseñada para trabajar con ese tipo de etiquetado, *one-hot*, mientras que sigmoid no. Esta tendencia se aprecia en ambas gráficas figuras 4.17 y 4.18. Destacar las métricas obtenidas por el modelo 16, este obtiene un valor de exactitud mediocre, del 60 %, siguiendo el patrón esperado, pero un valor-F1 superior al resto de modelos entrenados con el mismo conjunto de datos. Esto cuadra con lo visto en la gráfica figura 4.6, donde se aprecia que el modelo obtiene unos datos menores pero más equilibrados.

Se puede observar una ligera mejora en ambas métricas al usar datos multi-etiquetados y función de activación sigmoid, pero en la mayoría de los casos esta diferencia no es tan notable. De estos modelos el que mejor exactitud presenta es, con diferencia, el 23, obteniendo un 89 %, seguido por el 22 obteniendo un valor ligeramente menor.

Estas tendencias se ven replicadas en la gráfica figura 4.18, donde se aprecia que el valor-f1 sigue el mismo patrón pero en este caso las diferencias entre los modelos son más drásticas, es debido a esto por lo que el eje de ordenadas empieza en un valor de 0, a diferencia de figura 4.17. Esto hace de modelo 23 un modelo aún más superior al resto, puesto que obtiene un valor-F1 significativamente superior al resto, estando cerca del 80 %.

Destacar el buen rendimiento del modelo 22. Esto difiere con lo que se esperaría, ya que ha sido entrenado con menos datos que los posteriores a él. Esto, como ya se ha

explicado en apartado 4.1.2 es debido a que el modelo ha sido entrenado con una parte de los datos usados para realizar las pruebas, es por esto que las métricas están infladas y no representan del todo el rendimiento del modelo.

4.1.4. Conclusiones parciales

Como conclusión final extraída de los resultados de estas pruebas comentar que el hecho de usar técnicas de aumentado de datos para transformar un conjunto de datos de tipo *one-hot* en uno multi-etiquetado resultó en un impacto positivo en el rendimiento de los modelos, mientras que entrenar un modelo con función de activación sigmoid con datos no multi-etiquetados no da buenos resultados. Destacar también la importancia de separar los conjuntos de entrenamiento y de pruebas, ya que las métricas infladas resultantes pueden generar conclusiones erróneas y es una práctica poco recomendable y que hay que llevar a cabo con cautela. Finalmente destacar la importancia de entrenar varios modelos, al menos con conjuntos de datos y problemas pequeños, donde los tiempos de entrenamiento y los recursos necesarios no son muy altos ya que la aleatoriedad del entrenamiento y los pesos iniciales tiene un impacto alto, como se aprecia en las gráficas figuras 4.17 y 4.18, donde el modelo 23 obtiene unos resultados significativamente superiores al resto, aún siendo entrenado en las mismas condiciones que el resto y con los mismos datos que aquellos posteriores a él.

4.2. Textos seleccionados

Una parte esencial de la validación del rendimiento de un modelo es analizar las salidas generadas sobre datos seleccionados manualmente para analizar, de una manera más directa como este procesa las entradas y “piensa”. De esta manera se puede ver que palabras, frases y patrones hacen que el modelo tome ciertas decisiones, análisis que, como se explicará más adelante, puede resultar crucial en el último apartado de la validación de los modelos.

Se han elegido manualmente una serie de textos para clasificar con el modelo que mejores resultados obtuvo en las pruebas anteriores, el 23. De esta manera se puede apreciar de una manera directa, como este funciona y razona y se puede analizar si es capaz de interpretar correctamente los textos.

Estos textos son fragmentos del abstract de una serie de publicaciones científicas escogidas manualmente de Scopus tras una primera búsqueda relacionada con algún tema u objetivo en concreto. No se presentan textos relacionados con todos los objetivos ya que se consideró que no aportaría información relevante analizar aquí en profundidad la literatura referente a todos.

A continuación se presentan los textos seguidos de una tabla conteniendo los tres objetivos con un mayor porcentaje asignado. No se han incluido los asociados a más

objetivos porque llegada una determinada magnitud, el porcentaje de pertenencia deja de tener sentido.

Se han realizado estas búsquedas como base sobre la que extraer textos relevantes, relacionados con sostenibilidad. No se espera que sean clasificados acorde a ningún objetivo en concreto, únicamente en base a la búsqueda, si no que esto dependerá íntegramente del texto extraído.

4.2.1. Textos Sostenibilidad y sanidad

Estos textos han sido sacados de los abstracts resultantes al realizar la siguiente búsqueda en *Scopus*: sostenibilidad y sanidad

De entre todas las publicaciones resultantes de la búsqueda se eligieron 3 al azar. Teniendo en cuenta la búsqueda realizada se esperaría que el modelo los clasificase como relacionados con el objetivo 3.

Texto 1

“Background: Suriname is a uppermiddle-income country with a relatively high prevalence of preventable pregnancy complications. Access to and usage of high-quality maternity care services are lacking. The implementation of group care (GC) may yield maternal and child health improvements.”[46]

TABLA 4.3. PORCENTAJES TEXTO 1 - SANIDAD

ODS3	ODS17	ODS2
0.524	0.041	0.032

El texto extraído habla sobre la importancia del acceso a servicios sanitarios por parte de las mujeres embarazadas para mejorar la salud de las madres y los infantes.

Como se esperaba, el modelo le asigna un 52 % de pertenencia al objetivo 3 ya que se habla de sanidad. El siguiente objetivo con mas porcentaje asignado es el 17 con un 4 % y el 2 con un 3 %. Estos últimos pueden ignorarse ya que tienen un porcentaje asignado muy bajo, al igual que el resto de objetivos.

Texto 2

“Globally, forests serve as the largest storehouses for (non) indigenous trees, and are essential for the ecosystems’ sustenance, yet, increased deforestation practices associated with activities, such as tree logging, agriculture, and urban expansion continue to put pressure on existing forest areas, leading to massive land use cover change”[47]

TABLA 4.4. PORCENTAJES TEXTO 2 - SANIDAD

ODS15	ODS11	ODS13
0.997	0.047	0.041

Este texto habla sobre el peligro de la deforestación sobre los ecosistemas terrestres y como esta práctica pone en peligro las especies no indígenas.

Pese a ser resultado de una búsqueda sobre sanidad, este texto está altamente relacionado con el objetivo 15, el referente a los ecosistemas terrestres. Esto se ve reflejado en la predicción del modelo, asignándole a este objetivo un porcentaje del 99.7 %, mientras que al resto menos de un 5 %.

Texto 3

“With high staff vacancies in the health services, it is important to consider pragmatic methods of data collection for implementation evaluation. This paper presents a cross-sectional rapid evaluation of a handheld medical device designed for remote examinations, piloted in Northern England.”[48]

TABLA 4.5. PORCENTAJES TEXTO 3 - SANIDAD

ODS3	ODS17	ODS16
0.906	0.093	0.019

Este último texto sobre sanidad si está más relacionado con el tema principal, habla sobre un dispositivo médico diseñado para realizar examinaciones en remoto.

Como cabe esperar el objetivo 3 tiene asignado un 90 % de pertenencia. Adicionalmente le asigna cerca de un 10 % al objetivo 17, esto puede ser debido a que se habla de métodos para la recogida de datos, lo cual está más relacionado con este objetivo.

4.2.2. Textos Sostenibilidad e industria

En esta segunda búsqueda se centra el enfoque en la sostenibilidad y la industria.

No se espera ningún objetivo en concreto ya que hay muchos que pueden relacionarse con temas industriales, pero los principales son el 7, relacionado con energía, 8 relacionado con crecimiento económico, 9 relacionado de manera más directa con la industria, el 11 relacionado con las ciudades y finalmente el 12 relacionado con consumo y producción responsable.

Texto 1

“With the world’s population continuing to grow exponentially, with many ‘food deserts’ across the globe, including even in rich countries, food production is more important than ever. Finding alternative ways to produce food, in a sustainable way, is increasingly important and something that is on the minds of scientists, engineers, policy makers, and other professionals.”[49]

TABLA 4.6. PORCENTAJES TEXTO 1 - INDUSTRIA

ODS2	ODS12	ODS15
0.733	0.062	0.036

Este primer texto se escapa ciertamente del tema principal del apartado, la industria. Hablando de los llamados desiertos de comida y la importancia de la producción e industria alimenticia, y formas de hacerlas más sostenibles.

Leyendo el texto se ve la relación directa con el objetivo 2, debido a que la comida es su tema principal. También se encuentra el objetivo 12, aunque con un porcentaje bajo, del 6 %. Esto es menos de lo que se esperaría del texto ya que se habla de como conseguir una producción sostenible de comida. Esto puede ser resultado de la corta longitud del texto.

Texto 2

“With rising pollution emissions, it is vital to devise regulatory policies that ensure sustainable development. Green innovation offers an alternative strategy, fostering economic progress and environmental sustainability. While existing literature supports the positive role of green innovation in firm-level decisions, its specific impact on alleviating environmental regulation pressures remains unexplored.”[50]

TABLA 4.7. PORCENTAJES TEXTO 2 - INDUSTRIA

ODS13	ODS12	ODS9
0.164	0.0888	0.056

Este último texto del apartado de industria habla de la importancia de crear políticas regulatorias en referencia al aumento de emisiones. El principal texto predecido es el 13, como se esperaría hablando de este tema, pero este tiene asignado un porcentaje bajo. Esto puede ser, de nuevo, debido a la longitud del texto.

4.2.3. Textos Sostenibilidad e igualdad

En este apartado se han buscado textos relacionados con sostenibilidad e igualdad. Temas mas afines al objetivo 5 y 10.

Texto 1

“It is hard to establish whether a company supports internal sustainability efforts (ISEs) like gender equality, diversity, and general staff welfare, not least because of a lack of methodologies operationalizing these internal sustainability practices, and of data honestly documenting such efforts. We developed and validated a six-dimension framework reflecting Internal Sustainability Efforts (ISEs), gathered more than 350K employee reviews of 104 major companies across the whole US for the (2008-2020) years, and developed a deep-learning framework scoring these reviews in terms of the six ISEs.”[51]

TABLA 4.8. PORCENTAJES TEXTO 1 - IGUALDAD

ODS5	ODS12	ODS17
0.459	0.253	0.201

Este primer texto habla sobre las dificultades a la hora de esclarecer si una empresa está adoptando medidas a favor de la sostenibilidad, entre ellas igualdad de género y como su sistema desarrollado ayuda a determinarlo.

El principal objetivo clasificado por el modelo es el 5, con un 46 % asignado. Esto es lo que se podía esperar leyendo el texto ya que este tiene la igualdad de género como tema principal. También le asigna un 20 % a los objetivos 12 y 7, métricas que no se esperarían de este texto, más aún cuando la única mención posible relacionada es la que se hace a los esfuerzos por la sostenibilidad, pero esta se hace en un carácter cerrado como se menciona a continuación de la misma en el texto.

Texto 2

“Our aim was to evaluate how well we carried out authentic co-creation of an intervention to support midwives have a dialogue about alcohol consumption with pregnant women. Patient involvement: Recent maternity service users including women with experience of harm due to alcohol during pregnancy provided feedback on the design, conduct and dissemination of the study.”[52]

TABLA 4.9. PORCENTAJES TEXTO 2 - IGUALDAD

ODS5	ODS3	ODS2
0.424	0.287	0.076

Este texto habla de una iniciativa para concienciar a mujeres embarazadas del peligro que acarrea el consumo de alcohol durante el embarazo.

Los objetivos con más porcentaje asignado por el modelo son el 5 y el 3, resultados que cuadran con el texto dado. El 5 es de los dos el que menos podría cuadrar dado que no se menciona la igualdad de género, pero siempre que se mencionan mujeres este objetivo está presente. El único problema con los porcentajes asignados es la magnitud de los mismos, estos son más bajos de lo que se esperaría del modelo, lo cual sería un 50 % o más para ambos, mientras que los reales son un 42 % y un 28 %.

4.2.4. Textos Sostenibilidad y energía

Este último grupo de textos están relacionados con al sostenibilidad y la energía, dado esto se esperaría una mayor presencia del objetivo 7.

Texto 1

“Additive Manufacturing (AM) or 3D printing techniques use fused layers of the material to build cross sectional geometry of product. As variable processing parameters have an impact on the product quality, it is crucial to ascertain relationships of AM process parameters, productivity, sustainability, and structure performance.”[53]

TABLA 4.10. PORCENTAJES TEXTO 1 - ENERGÍA

'ODS12'	'ODS9'	'ODS8'
0.713	0.473	0.102

Este primer texto habla de un tipo de manufacturación aditiva y los impactos en sostenibilidad, productividad y rendimiento que tiene.

Las predicciones de este modelo son lo que se podría esperar de un texto así, una mayor presencia del objetivo 12, con un 71 % de pertenencia ya que se habla de sostenibilidad en un entorno industrial y de consumo y el 9, con un 47 %, que, pese a quedarse por debajo del umbral, al ser la diferencia de 3 puntos porcentuales se puede considerar como asignado para motivos de análisis. La asignación de este último objetivo resulta esperable y adecuada estando este relacionado con la industria y la innovación en la misma.

Texto 2

“Electric Vehicles are a suitable solution for sustainability in transportation applications. The era of wireless power transfer in free space began with the evolution of Tesla coils being energized by microwaves. High-frequency inverters act as the electrifier for high power wireless charging. DC-DC converters play an indispensable role in

converting the AC power from a high-frequency inverter to DC power to the battery or the energy storage system in the vehicle.”[54]

TABLA 4.11. PORCENTAJES TEXTO 2 - ENERGÍA

ODS9	ODS7	ODS11
0.261	0.147	0.090

Este texto habla sobre la transferencia inalámbrica de potencia eléctrica y del papel de los coches eléctricos como solución sostenible a los problemas de transporte.

Los porcentajes asignados a este texto son muy bajas, probablemente debido a la cantidad de tecnicismos usados y a que estos no están directamente relacionados con el vocabulario que se esperaría en un contexto de sostenibilidad. De todas formas los objetivos con mayor porcentaje si que son los que se espera de un texto como este, siendo el 9, con un 26 % y el 7 con un 15 %, estando el primero relacionado con la industria y el segundo con la energía eléctrica.

Texto 3

“Aquaculture is a form of agriculture that is practised in the Sistan region. This practice not only enhances water usage efficiency but also generates additional income for the local farmers. The extensive dependence of the aquaculture sector on energy and chemicals poses a jeopardises environmental stability and endangers the sustainability of production in the long run.”[55]

TABLA 4.12. PORCENTAJES TEXTO 3 - ENERGÍA

ODS6	ODS12	ODS7
0.315	0.251	0.230

Este último texto habla sobre la práctica denominada *Aquaculture* y como esta afecta al consumo de agua y a la sostenibilidad.

Este texto es de los más ambiguos, de todas formas si se analizan los objetivos prede- cidos estos tienen bastante sentido, siendo el que mayor porcentaje ha obtenido el 6, con un 32 %, seguido del 12 con un 25 % y el 7 con un 23 %. Estos están correctamente asig- nados, estando el primero relacionado con el consumo de agua, el segundo con consumo y producción sostenible y el último con la energía.

4.2.5. Conclusiones parciales

Como conclusión de estos análisis destacar la importancia de la aparición de ciertas palabras en el texto, como por ejemplo *mujer* a la hora de clasificar el objetivo 5.

Adicionalmente destacar la excelente capacidad del modelo a la hora de identificar aparte del tema principal de un texto, temas secundarios que son mencionados de manera superficial.

4.3. Literatura científica

La ultima validación de los modelos se realizó usando un corpus de literatura científica, en concreto los abstracts de publicaciones realizadas en WoS. Dichos abstracts fueron extraídos por medio de unas consultas publicadas en un trabajo de la universidad de Estocolmo [56]. En dicho trabajo los autores proporcionan 17 de estas consultas especializadas, cada una en un objetivo, para resultar en publicaciones relacionadas con cada uno en concreto. La fiabilidad de estas métricas se discutirá más adelante pero sirven como una buena base para analizar el rendimiento de los modelos en un entorno similar al final.

Se extrajeron un total de 1000 publicaciones de cada objetivo, siendo este el máximo permitido por la propia plataforma, destacar que no todas las publicaciones vienen con abstract incluido por lo que el número de textos clasificados es menor de 1000. Posteriormente los abstracts de dichas publicaciones se clasificaron usando el modelo y se contó el numero total de objetivos asignados a cada consulta. Se espera que haya cerca de 1000 objetivos asignados correctamente por consulta.

4.3.1. Modelos de redes recurrentes

De todos los modelos generados se muestran los resultados de los más relevantes, cuadro 4.14, esto se calculó realizando la media de etiquetas asignadas entre los 17 objetivos por cada modelo, representadas en cuadro 4.13, y seleccionando los 4 mejores, estos coinciden con los 4 primeros.

TABLA 4.13. MEDIA DE CLASIFICACIONES POR MODELO RECURRENTE

Modelo	Media de clasificaciones
1	165.82
2	139.94
3	149.76
4	134.29
5	100.88
6	113.29
7	108.70
8	85.94
9	97.58
10	74.58

En la tabla, cuadro 4.14 se muestra el número de objetivos asignados por cada modelo

basado en una red recurrente a todos los textos científicos extraídos, en ella se puede apreciar que el número es relativamente bajo para todos, siendo el que mayor rendimiento muestra el modelo 1, aunque esto no se ve reflejado en todos los objetivos. Demostrando así la poca fiabilidad de estos modelos a la hora de realizar un estudio cuantitativo ya que, aunque no se espere que se clasifiquen los 1000 textos de manera correcta, se obtienen unos resultados muy pobres, clasificando, en el mejor de los casos, 576 textos y de media 117 por objetivo, estando muy lejos de los 1000.

TABLA 4.14. CLASIFICACIONES DE CADA MODELO
RECURRENTE POR OBJETIVO

Objetivo	Modelo 1	Modelo2	Modelo3	Modelo4	ODSs Totales
ODS1	576	416	294	381	855
ODS2	59	140	98	143	913
ODS3	128	110	90	22	284
ODS4	56	41	21	35	145
ODS5	345	256	315	154	619
ODS6	169	78	63	211	851
ODS7	24	129	43	256	788
ODS8	31	42	5	93	975
ODS9	31	178	31	40	965
ODS10	81	54	64	67	831
ODS11	328	233	175	341	987
ODS12	231	16	67	173	830
ODS13	99	134	566	21	914
ODS14	18	115	171	58	983
ODS15	169	224	320	62	978
ODS16	247	44	196	173	777
ODS17	227	169	27	53	994

4.3.2. Modelos de BERT

A continuación, cuadros 4.15 a 4.18 se muestran las clasificaciones realizadas por los modelos basados en BERT, de estos se espera un mejor rendimiento ya que son más capaces a la hora de entender los textos.

Modelos softmax

En primera instancia, cuadro 4.15, se muestran las clasificaciones de los modelos con función de activación softmax, entrenados con los mismos datos que los modelos basados en redes recurrentes, cuadro 4.14, se puede observar que, en la mayoría de los casos se obtienen unos resultados muy superiores, con unas notables excepciones, presentes en todos los modelos. Estas son en los objetivos del 8 al 17. En estos últimos se clasifican correctamente muy pocos textos, siendo en todos los casos menos de la mitad de los textos

totales. Esto puede ser debido a un problema en las consultas, tema abordado al final del apartado, o un problema con los modelos. Si resulta ser un problema con las consultas, este patrón se apreciará en el resto de tablas, si resulta ser un problema con los modelos este patrón solo estará presente en esta tabla.

TABLA 4.15. CLASIFICACIONES MODELOS SOFTMAX

Objetivo	Modelo 11	Modelo13	Modelo14	Modelo15	ODSsTotales
ODS1	742	670	737	693	855
ODS2	796	808	795	786	913
ODS3	261	269	258	277	284
ODS4	127	126	118	127	145
ODS5	573	566	580	528	619
ODS6	461	457	504	479	851
ODS7	543	601	741	649	788
ODS8	52	99	39	39	975
ODS9	158	229	105	330	965
ODS10	74	107	55	252	831
ODS11	93	166	179	248	987
ODS12	126	101	145	144	830
ODS13	56	65	81	65	914
ODS14	189	131	91	323	983
ODS15	414	376	383	459	978
ODS16	388	350	422	561	777
ODS17	292	161	162	285	994

Modelos sigmoid

Las siguientes clasificaciones son las de los modelos con función de activación sigmoid pero entrenados con los datos iniciales, los *one-hot*, como se puede observar en la tabla cuadro 4.16 estos no obtienen unos resultados buenos a excepción del modelo 16 que obtiene un número de clasificaciones excelente. De todas formas, y como se analizó en la gráfica figura 4.6, este modelo, aún obteniendo una buena métrica de exhaustividad, tiende a asignar muchos objetivos pero con menos acierto.

Adicionalmente la tendencia observada en los datos anteriores, cuadro 4.15, sigue presente, aunque se aprecia menos ya que dos de los modelos clasifican muy pocos objetivos y el otro restante clasifica de más. De todas formas se sigue observando una reducción en el número de objetivos clasificados a partir del 8.

Entrenado con datos one-hot

TABLA 4.16. CLASIFICACIONES MODELOS SIGMOID - DATOS INICIALES

Objetivo	Modelo 16	Modelo17	Modelo18	ODSs Totales
ODS1	845	321	287	855
ODS2	902	55	213	913
ODS3	284	71	104	284
ODS4	144	34	88	145
ODS5	615	162	317	619
ODS6	846	236	235	851
ODS7	755	176	487	788
ODS8	417	1	7	975
ODS9	903	0	78	965
ODS10	562	0	11	831
ODS11	422	5	6	987
ODS12	740	0	17	830
ODS13	772	1	3	914
ODS14	954	2	47	983
ODS15	625	69	345	978
ODS16	582	143	170	777
ODS17	859	1	47	994

Entrenado con datos aumentados

A partir de este punto se muestran los datos obtenidos por los modelos entrenados con datos multi-etiqueta, siendo estos primeros aquellos entrenados con los datos iniciales aumentados.

En la tabla cuadro 4.17 se puede apreciar un buen rendimiento general de todos los modelos, destacando el obtenido por el modelo 21, ya que este obtiene un buen numero de clasificaciones para todos los objetivos. Esto contrasta con el resto de métricas analizadas, figuras 4.17 y 4.18, en las que este modelo no obtiene unos resultados destacables. La explicación de esta mejora no está clara y pueden ser causada por varios factores. Destacar también que las clasificaciones de los tres modelos para todos los objetivos anteriores al 8 son bastante similares, destacando ligeramente el rendimiento del modelo 21, pero en una medida que cabría esperar, sobre todo siendo su valor-F1 ligeramente superior que el del resto.

Adicionalmente indicar que le patrón analizado de una disminución en el numero de clasificaciones a partir del objetivo 8 sigue estando presente en estas métricas, obteniendo estos objetivos un número de clasificaciones cercano a la mitad de los esperado en la mayoría de los casos.

TABLA 4.17. CLASIFICACIONES MODELOS SIGMOID - DATOS AUMENTADOS

Objetivo	Modelo 19	Modelo20	Modelo21	ODSsTotales
ODS1	536	465	705	855
ODS2	797	849	859	913
ODS3	247	257	280	284
ODS4	138	135	141	145
ODS5	579	605	608	619
ODS6	529	593	509	851
ODS7	711	638	631	788
ODS8	26	139	96	975
ODS9	291	198	497	965
ODS10	137	60	377	831
ODS11	172	199	628	987
ODS12	175	409	348	830
ODS13	317	486	659	914
ODS14	229	204	531	983
ODS15	448	444	420	978
ODS16	506	540	536	777
ODS17	322	435	357	994

Entrenado con datos aumentados extraídos de internet

Finalmente se presentan las clasificaciones generadas por los últimos modelos, aquellos entrenados con los datos finales, a excepción del modelo 22 que fue entrenado únicamente con los datos multi-etiqueta extraídos de internet.

Estos modelos, como se presenta en la tabla, cuadro 4.18, y a diferencia de lo que cabría esperar y de lo que indica las métricas de las gráficas, figuras 4.17 y 4.18, presentan un rendimiento menor que los modelos anteriores, cuadro 4.17.

De todos los modelos, es el 23 el que obtiene los mejores resultados, presentando, en este caso, una diferencia notable con el modelo 22 confirmando así que las métricas de este último, representadas en las gráficas, figuras 4.17 y 4.18, estaban infladas debido a que se entrenó con parte de los datos presentes en el conjunto de pruebas usado para calcularlas.

Destacar finalmente la reiteración del patrón identificado en el resto de tablas anteriores, obteniendo un número de clasificaciones menor en todos los objetivos a partir del 8. En este caso es ciertamente más sutil, con casos como el objetivo 16, en el que se clasifican casi tantos textos como en el 7. Y el objetivo 15, obteniendo una fracción menor de textos clasificados pero aún así es ciertamente superior al resto de objetivos del patrón.

TABLA 4.18. CLASIFICACIONES MODELOS SIGMOID - DATOS EXTRAÍDOS DE INTERNET/FINALES

Objetivo	Modelo 22	Modelo23	Modelo24	Modelo25	Modelo26	ODSsTotales
ODS1	180	765	208	139	232	855
ODS2	408	811	342	545	781	913
ODS3	194	275	119	234	233	284
ODS4	11	105	8	52	36	145
ODS5	430	603	347	512	364	619
ODS6	379	436	238	381	499	851
ODS7	304	442	174	249	277	788
ODS8	21	112	57	82	129	975
ODS9	31	184	33	59	160	965
ODS10	4	31	8	14	48	831
ODS11	1	202	1	7	8	987
ODS12	165	125	229	130	314	830
ODS13	241	360	195	236	281	914
ODS14	59	118	66	56	72	983
ODS15	366	411	342	390	380	978
ODS16	174	407	100	414	266	777
ODS17	313	298	326	392	452	994

4.3.3. Análisis de los datos

Como se ha explicado en el apartado anterior, se ha identificado un patrón en las clasificaciones, a partir del objetivo 8 se identifican significativamente menos textos asignados a cada objetivo. Esto podría ser problema de los modelos o de los datos usados para el entrenamiento, de todas formas esta posibilidad pasa a un segundo plano ya que la presencia del mismo patrón, en mayor o menor medida, independientemente de la arquitectura del modelo y de los datos usados para su entrenamiento hace de esto una posibilidad remota. Es por esto por lo que se ha decidido analizar los datos extraídos haciendo uso de la herramienta *VosViewer* [57]. Esta herramienta, a partir de un corpus de artículos científicos exportados de bases de datos científicas como WoS, genera una gráfica de agrupación con todos los términos con mayor presencia entre los documentos, junto con las relaciones y magnitudes de los mismos. De esta manera se puede analizar la taxonomía de un corpus científico de manera rápida y sencilla.

A continuación se muestran los mapas de dispersión generados para todos los objetivos a partir del 8, con el fin de analizar la taxonomía presente en cada uno y esclarecer el porque del patrón. Si en estos mapas se aprecia una taxonomía diferente a la que se esperaría para cada objetivo, serviría como justificación para la aparición de este tipo de patrón. Adicionalmente se puede analizar que otros objetivos se clasifican cuando se le presentan estos textos a los modelos como validación final de los mismos ya que si clasifican los objetivos relacionados con la taxonomía extraída se demostrará el correcto funcionamiento de los mismos.

Análisis objetivo 8

Esta primera gráfica, figura 4.19, muestra las ocurrencias y relaciones de palabras, términos y expresiones presentes en los textos relacionados con el objetivo 8. Estando este objetivo relacionado con el crecimiento económico y el trabajo decente se esperaría encontrar en mayor o menor magnitud, términos económicos y relacionados con el trabajo.

Analizando el mapa generado por VosViewer se puede ver que los términos más cunes son *effect, access, management* y *health*. Siendo estos los términos principales se puede entender que los modelos no fueran capaces de clasificar correctamente los textos ya que son términos muy generales y poco relacionados con el objetivo 8. Analizando más en profundidad la figura, se puede apreciar algún término relacionado con la economía y el trabajo como puede ser *income* o *employment*. Adicionalmente analizado las agrupaciones generadas por el programa se identifican cinco grupos diferentes, ninguno de los cuales directamente relacionado con el objetivo 8:

- **Grupo amarillo:** Este primer grupo está poco relacionado con los temas principales del objetivo 8, estando más directamente relacionado con la energía. De todas formas, estos son términos que podrían estar, en ocasiones, ligeramente relacionados con el crecimiento económico, siendo la energía parte esencial del mismo.
- **Grupo azul:** Este grupo es de todos el menos relacionado, incluyendo términos relacionados con la salud. Al igual que en el caso anterior son términos que podrían encontrarse a la par de temas económicos pero en ningún caso se esperaría encontrarlos en un primer plano.
- **Grupo rojo:** Este grupo tiene una relación algo más directa, siendo, a su vez, de carácter más general. Incluyendo términos de gestión, planificación y metodología, lo cual no choca con lo que se esperaría si se hablara de economía y trabajo.
- **Grupo verde:** Este es de todos, el grupo más grande, conteniendo como término principal *effect*. La alta presencia de este término es difícil de interpretar, pudiendo hacer referencia a estudios que analizan el efecto de medidas o investigaciones. Si se analizan las conexiones se puede ver que las mayorías son entre este grupo y el amarillo, pudiendo esto indicar un alto número de trabajos investigando el efecto de avances o políticas en el ámbito energético. Adicionalmente la presencia de términos como *income* o *gdp* puede indicar una mayor relación entre este grupo y el objetivo 8. De todas formas la presencia de estos términos en menor medida y relacionados con otros altamente especializados, como son los del grupo amarillo pude relevar al objetivo 8 a un segundo plano en cuanto a la clasificación.
- **Grupo morado:** Este último y menor grupo se encuentra en una posición más central en el diagrama indicando su presencia más general. Este está altamente relacionado con los recursos naturales y la comida. Términos que, como ya se ha visto

con otros grupos, no extrañaría verlos de la mano de otros hablando de crecimiento económico.

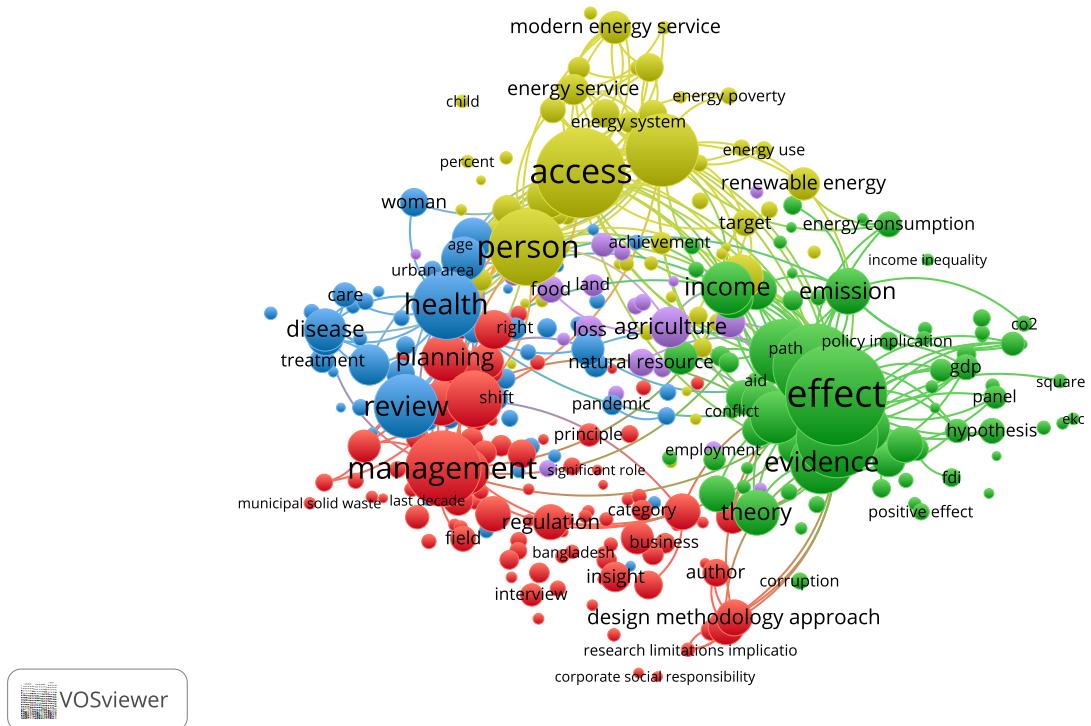


Fig. 4.19. Mapa de relaciones del objetivo 8

En general este diagrama, figura 4.19, confirma que la taxonomía de este conjunto está altamente dividida, de hecho se identifican términos relacionados con más objetivos, como los son el 2 (relacionado con la alimentación), el 7 (relacionado con la energía), el 3 (relacionado con la salud) e incluso el 17 (relacionado con la gestión y las regulaciones). Es por esto por lo que se decidió realizar un análisis más profundo.

Debido a la heterogeneidad identificada en los términos, se decidió analizar el número de clasificaciones totales generadas por el modelo 23 sobre el conjunto de datos del objetivo 8. Métricas representadas en la tabla, cuadro 4.19. Como se puede apreciar en esta, el número de clasificaciones está muy repartido entre varios objetivos, destacando los objetivos, 3, 7, 8, 9, 12 y 17, obteniendo más de 100 clasificaciones. Este patrón encaja casi a la perfección con las conclusiones extraídas del mapa generado por VosViewer, figura 4.19, en el que se identificaron todos estos objetivos, a excepción del 9 y el 12, en la taxonomía. Destacar también la presencia, aunque en menor medida, del objetivo 2, obteniendo 93 clasificaciones, lo cual coincide con la presencia del quinto grupo identificado en la figura 4.19.

TABLA 4.19. CLASIFICACIONES ASBTRACTS ODS8 - MODELO

23

Objetivo	Clasificaciones
ODS1	45
ODS2	93
ODS3	126
ODS4	18
ODS5	23
ODS6	33
ODS7	173
ODS8	112
ODS9	137
ODS10	21
ODS11	75
ODS12	121
ODS13	81
ODS14	19
ODS15	63
ODS16	12
ODS17	127

Análisis objetivo 9

A continuación se muestra, en al ráfica, figura 4.20, el diagrama relacionado con el objetivo 9, este obtuvo, en la clasificaciones del objetivo 23, cuadro 4.18, un resultado ligeramente superior al del objetivo 8, por lo que se esperaría encontrar una taxonomía mas relacionada con el objetivo 9, contenidoendo términos relacionados con la industria, innovación e infraestructura.

Analizando la figura figura 4.20 se pueden identificar unos términos principales como *relationship, literature, affordable housing y building* divididos en cuatro grupos:

- **Grupo rojo:** este grupo incluye términos relacionados con las corporaciones y relaciones. De el diagrama no se puede extraer un tema concreto de este grupo.
- **Grupo verde:** este grupo habla de temas como las cadenas de producción y en un primer plano, el termino “literatura”. Esto resulta ciertamente extraño ya que es un término que, aunque pudiendo tener un papel central en cuanto a número de ocurrencias, no está directamente relacionado con ningún objetivo.
- **Grupo azul:** este es uno de los dos grupos más claros y con un tema principal más identifiable, en este caso este tema es la vivienda, en concreto la vivienda asequible, la construcción y la infraestructura. Temas más directamente relacionados con el objetivo 9 aunque la se podría argumentar que la vivienda está más relacionada con el objetivo 11.

- **Grupo amarillo:** este último grupo es el más pequeño de todos, contando con términos sobre energía, construcciones y viviendas pero desde un punto de vista diferente al azul. Este también se podría identificar como relacionado con el objetivo 9 pero al igual que el azul también se podría argumentar que esta relación es más directa con otros objetivos como el 7 o el 11.

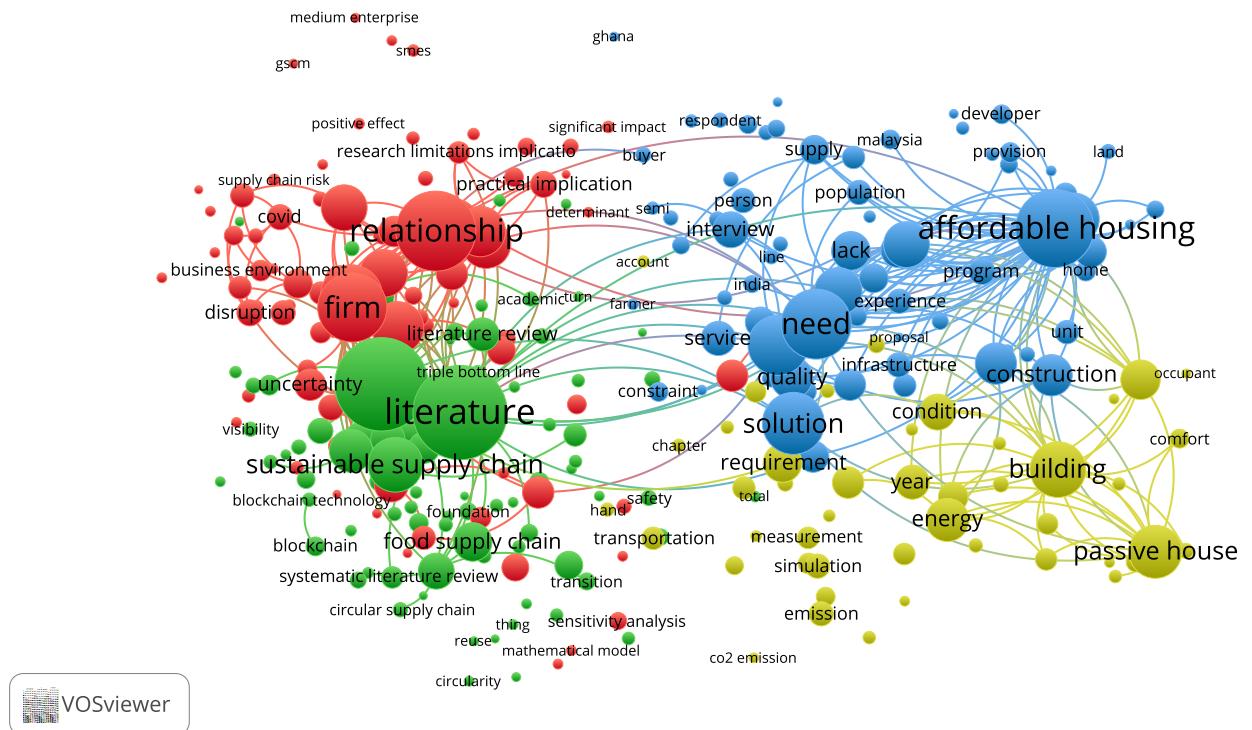


Fig. 4.20. Mapa de relaciones del objetivo 9

Como análisis final de este objetivo se analizaron las clasificaciones generadas por el modelo 23, de igual manera que se hizo con el objetivo 8. Estas, representadas en la tabla cuadro 4.20, muestran unos objetivos principales que son el 12 y el 17 pudiendo estar relacionados con los grupos verde y rojo respectivamente ya que los términos presentes en estos están más relacionados con estos dos objetivos que con el resto. También destacar las clasificaciones asignadas a los objetivos 9 y 11, esto coincide con lo analizado en los grupos amarillo y azul, en los cuales se identificaron términos relacionados con estos objetivos. También mencionar que la presencia de términos energéticos en el grupo amarillo se ve reflejada en las clasificaciones del objetivo 7.

TABLA 4.20. CLASIFICACIONES ASBTRACTS ODS9 - MODELO

23

Objetivo	Clasificaciones
ODS1	6
ODS2	46
ODS3	27
ODS4	0
ODS5	2
ODS6	5
ODS7	84
ODS8	27
ODS9	184
ODS10	2
ODS11	140
ODS12	302
ODS13	42
ODS14	4
ODS15	13
ODS16	5
ODS17	215

Análisis objetivo 10

Este objetivo es el más importante de analizar ya que es el que menos clasificaciones tiene asignadas por el modelo 23, un total de 31, cuadro 4.18. Este objetivo tiene como tema principal la reducción de las desigualdades por lo que se esperaría encontrar términos sociales.

Como se puede apreciar en el diagrama, figura 4.21, hay multitud de términos relacionados, divididos en cuatro grupos aunque se encuentran distribuidos de una manera más homogénea, demostrando que no hay tantos temas diferenciados. Estos 4 grupos principales son los siguientes:

- **Grupo verde:** Este primer grupo es de los más extensos, incluyendo términos como *China, export* y *tariff* indicando un tema muy diferente al objetivo 10, estando este más relacionado con las relaciones internacionales, tema principal del objetivo 17 e incluyendo términos como industria y agricultura, propios de los objetivos 8 y 2 respectivamente.
- **Grupo rojo:** este grupo tiene una estructura un tanto caótica, estando poco aglomerado, las ocurrencias de los términos de este grupo son menores que las del resto y, al no tener un tema en concreto, no se espera que tenga un impacto muy relevante en las clasificaciones.
- **Grupo amarillo:** este es el segundo grupo poco aglomerado, estando en su mayoría

entrelazado con el rojo. Presenta términos más relevantes y menos generales como lo son *countries*, *dispute* y *participation*. Estos términos tampoco están muy relacionados con el objetivo 10, haciendo referencia a temas más diplomáticos propios de otros objetivos, principalmente el 17.

- **Grupo azul:** este grupo final tiene un elemento central *access* relacionado con otros términos como *protección* o *intellectual property*. El tema principal de este grupo no está tan claro ya que no hay ningún término principal que lo aclare, si no uno mayoritario que conecta multitud de términos diferentes.

Como da a entender esta figura, la taxonomía presente en los datos del objetivo 10 no incluyen términos relacionados con igualdad ni con acciones sociales. Es por esto por lo que los modelos no clasifican ninguno de los textos incluidos en este conjunto como relacionados con el objetivo 10.

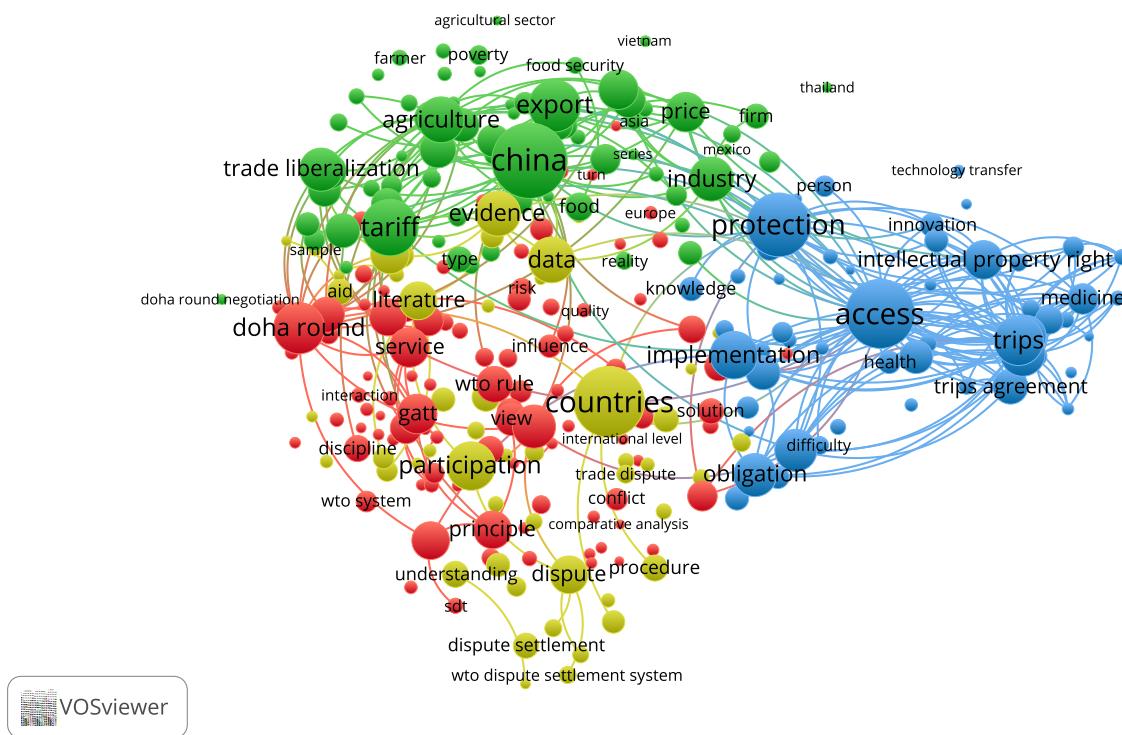


Fig. 4.21. Mapa de relaciones del objetivo 10

Como análisis adicional y para confirmar el análisis de los términos revelados en la figura 4.21, se incluyen en la tabla, cuadro 4.21, las clasificaciones totales generadas por el modelo 23 sobre los datos del objetivo 10. En esta se aprecia como solo 31 textos son clasificados como relacionados con el objetivo 10, mientras que al rededor de 100 son asignados a los modelos 2 y 8 estando estos relacionados en mayor o menor medida con los términos presentes en el grupo verde. Finalmente 564 son clasificados como relacionados con el objetivo 17, esto concuerda con gran parte de la taxonomía identificada.

TABLA 4.21. CLASIFICACIONES ASBTRACTS ODS10 - MODELO

23

Objetivo	Clasificaciones
ODS1	10
ODS2	115
ODS3	81
ODS4	2
ODS5	1
ODS6	3
ODS7	4
ODS8	111
ODS9	83
ODS10	31
ODS11	1
ODS12	21
ODS13	18
ODS14	5
ODS15	10
ODS16	12
ODS17	564

Análisis objetivo 11

A continuación se analizarán los datos asignados con el objetivo 11, este es un objetivo curioso ya que, como se aprecia en la tabla cuadro 4.18, el modelo 23 asigna más de 200 textos a este objetivo mientras que el resto no asigna más que 8 en el mejor de los casos. Haciendo referencia a la misma tabla, la cual contiene los mejores modelos conseguidos.

Analizando la figura figura 4.22 se puede observar una taxonomía con tres términos principales *digestion*, *scenario* y *municipal solid waste management*, estas, a priori, si están ciertamente relacionadas con el objetivo 11, igual no de una forma común pero es indudable que una buena gestión de los residuos es una tarea esencial si se desea conseguir ciudades y comunidades sostenibles, tema principal del objetivo 11. Adicionalmente se identifican tres grupos principales y un cuarto minoritario:

- **Grupo verde:** Este grupo contiene términos relacionados con la recolección, transporte y gestión de residuos. Términos relacionados principalmente con el objetivo 12 pero, en menor medida, relacionados con el 11.
- **Grupo azul:** La taxonomía presente en este grupo está definida por términos como sistema gestor de residuos, reciclaje y calentamiento global. Este también está relacionado con el objetivo 12 pero de igual manera relacionado con el 11.
- **Grupo rojo:** Este es el último grupo principal, en él se definen términos como digestión, compost, salud, comida y biomasa. Términos altamente relacionados con

el resto de grupos, principalmente la gestión de residuos, y al igual que el resto relacionados con el objetivo 12 y en menor medida con el 11.

- **Grupo amarillo:** Este último grupo está presente en mucha menor medida que el resto, analizarlo con la resolución generada no es posible ya que no se pueden identificar correctamente los términos.

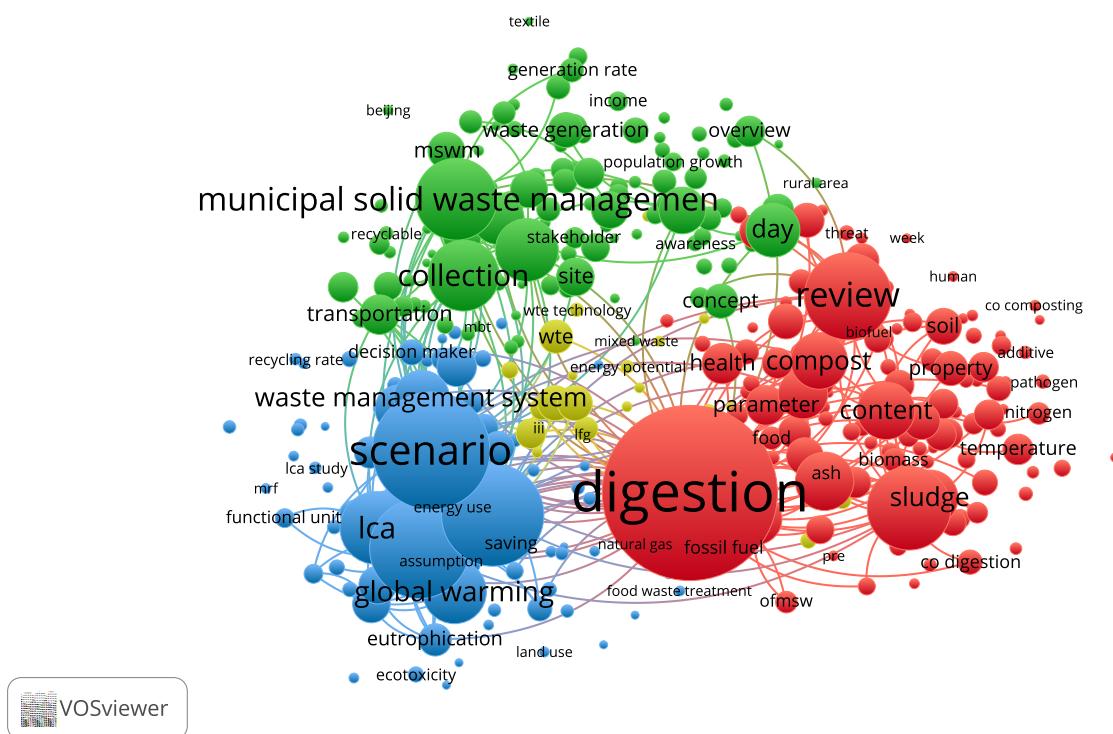


Fig. 4.22. Mapa de relaciones del objetivo 11

Para confirmar las conclusiones extraídas de la figura 4.22, se han incluido las clasificaciones totales del modelo 23 sobre todos los textos del objetivo 11, cuadro 4.22. En esta tabla se aprecia como casi 800 textos han sido clasificados como relacionados con el objetivo 12, mientras que 202 con el 11. Esto aunque dentro de lo que se esperaba, no se esperaba una diferencia tan grande. Destacar el numero de clasificaciones asignadas al objetivo 7, al rededor de 160, en este caso, en la taxonomía no se puede apreciar ningún termino relacionado con la energía, posiblemente sea fruto de múltiples publicaciones sobre generación de energía por medio de los residuos, pero esto es únicamente especulación.

TABLA 4.22. CLASIFICACIONES ASBTRACTS ODS11 - MODELO
23

Objetivo	Clasificaciones
ODS1	0
ODS2	27
ODS3	48
ODS4	2
ODS5	0
ODS6	28
ODS7	164
ODS8	5
ODS9	10
ODS10	1
ODS11	202
ODS12	782
ODS13	35
ODS14	5
ODS15	9
ODS16	0
ODS17	21

Análisis objetivo 12

El caso de este objetivo es el contrario al anterior, el 11. En este el modelo 23 asigna este objetivo a menos textos que el resto de modelos. En este caso la gráfica, figura 4.23, indica que la taxonomía de los datos de este modelo está altamente relacionada con temas de investigación, siendo la mayoría de los términos principales relacionados con este tema. A pesar de estar divididos en cuatro grupos, los temas de todos son similares.

De esta taxonomía no se puede sacar conclusiones claras, no teniendo términos significativos relacionados con ningún objetivo. La única explicación posible es que las publicaciones extraídas sean de temas muy diversos y no haya ningún tema principal entre ellas.

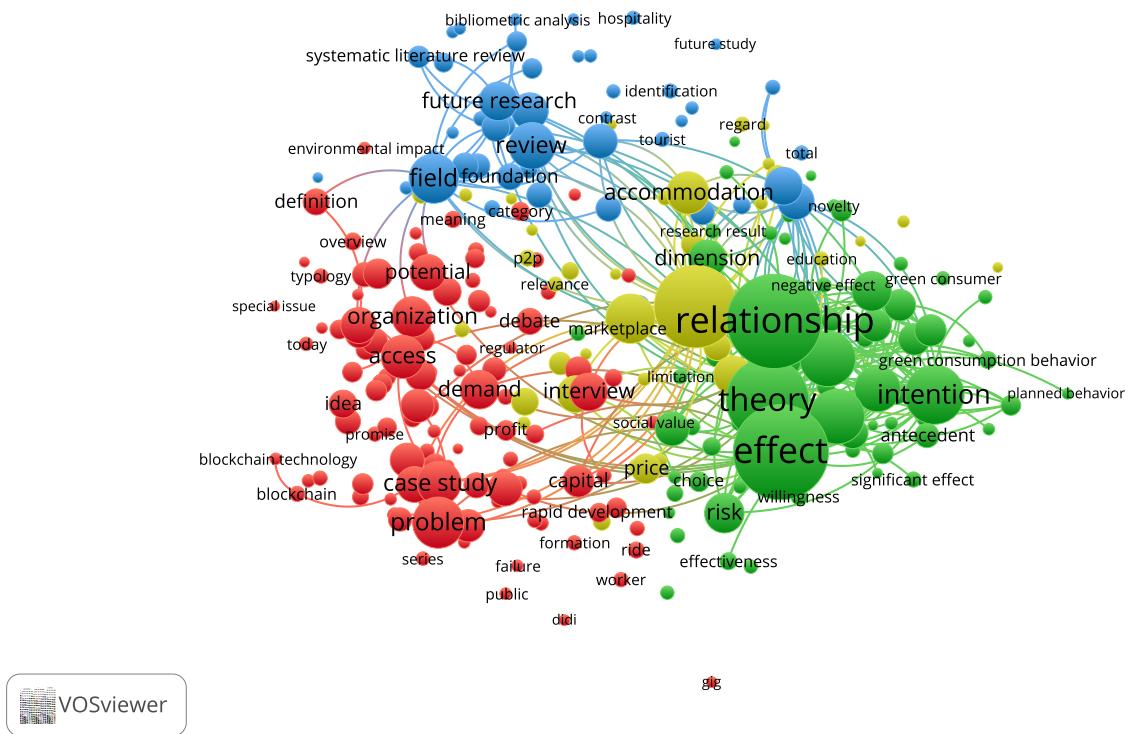


Fig. 4.23. Mapa de relaciones del objetivo 12

En la tabla cuadro 4.23 se aprecian todas las clasificaciones asignadas a los datos del objetivo 12. En estas se aprecia que se clasifican sobre todo los objetivos 8, 9, 12 y mayoritariamente el 17. En este caso no se identifica ninguna relación significativa entre la taxonomía extraída de los datos y las clasificaciones ya que esta primera es altamente ambigua y no aporta casi información relacionada con los objetivos.

TABLA 4.23. CLASIFICACIONES ASBTRACTS ODS12 - MODELO

23

Objetivo	Clasificaciones
ODS1	2
ODS2	9
ODS3	8
ODS4	2
ODS5	11
ODS6	2
ODS7	14
ODS8	224
ODS9	189
ODS10	3
ODS11	40
ODS12	125
ODS13	4
ODS14	0
ODS15	8
ODS16	2
ODS17	311

Análisis objetivo 13

Este objetivo consigue un número de clasificaciones mayor que los abordados hasta ahora. Es por esto que no se ha considerado necesario realizar un análisis tan en profundidad. Analizando la figura 4.24 se aprecia que uno de los 3 grupos, y a su vez el más grande en cuanto a magnitud, está directamente relacionado con este objetivo, incluyendo términos como cambio climático e impacto. Adicionalmente la existencia de otros dos grupos minoritarios con temas diferentes como lo son la vegetación, o los depósitos y sedimentación explica que se hayan clasificado un número tan relativamente bajo de textos como relacionados con el objetivo 13.

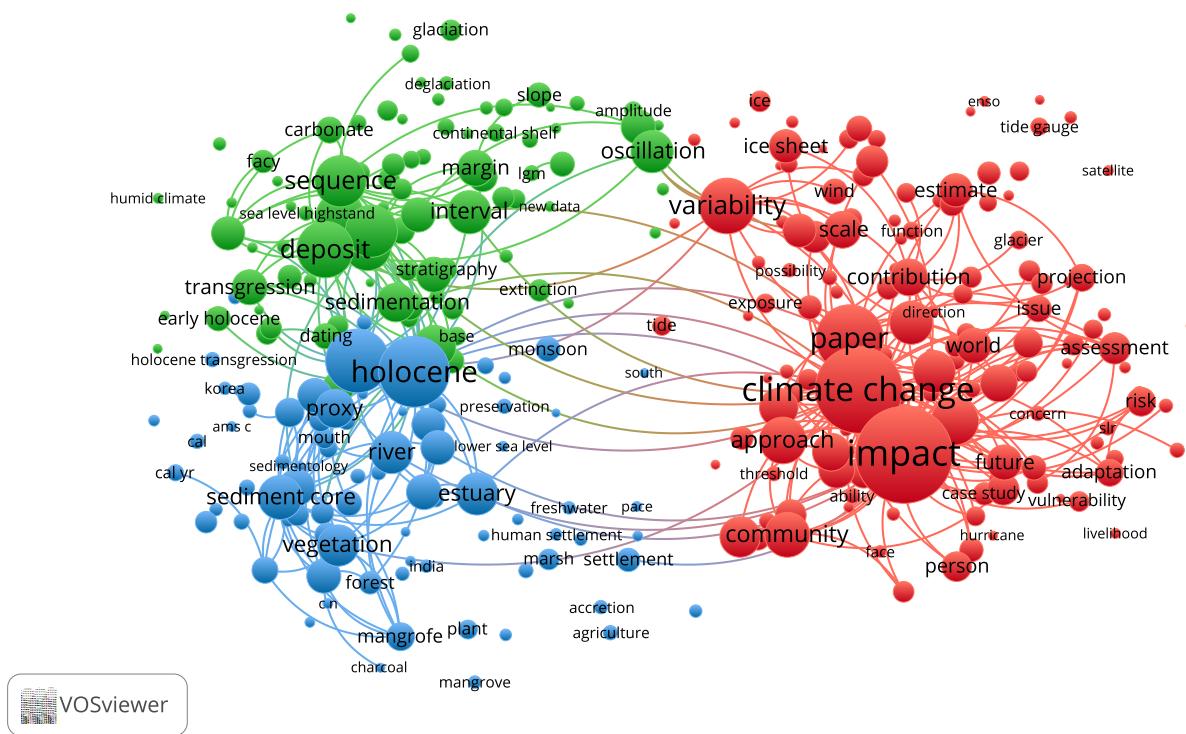


Fig. 4.24. Mapa de relaciones del objetivo 13

Análisis objetivo 14

Estos son el último de los objetivos que tiene una clasificación baja, 118 textos asignados por el modelo 23, cuadro 4.18. Esto contrasta con lo visto en la figura 4.25 ya que en esta se aprecia como la mayoría de los términos están relacionados con el agua, aún divididos en cuatro grupos, todos ellos están relacionados con este tema.

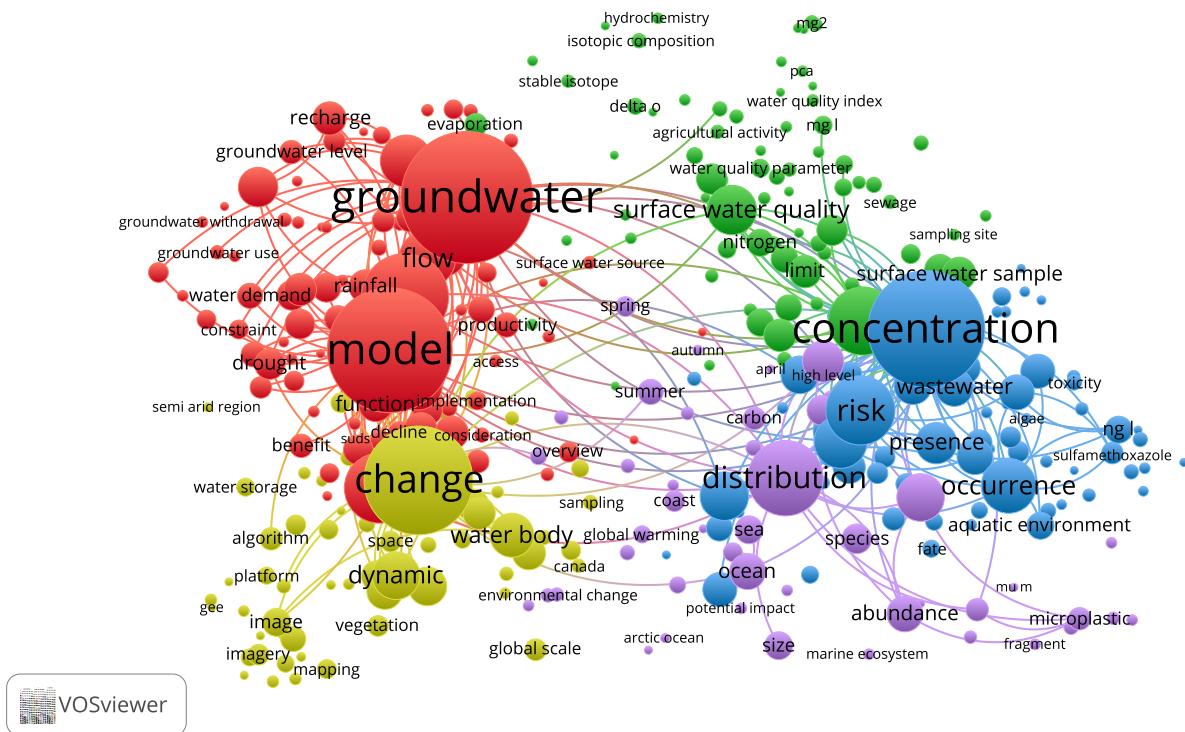


Fig. 4.25. Mapa de relaciones del objetivo 14

Como explicación se han incluido, cuadro 4.24, las clasificaciones de todos los textos relacionados con el objetivo 11. Es esta tabla se aprecia como la mayoría de los textos han sido clasificados como relacionados con el objetivo 6, aquel que habla sobre el la calidad y saneamiento del agua. Viendo esto y analizando de nuevo la figura 4.25 se puede entender el porque de este numero de clasificaciones tan bajo, ya que la taxonomía extraída no hace referencia a los ecosistemas ni ningún tipo de vida acuática por lo que está mas relacionado con el objetivo 6 que con el 11, estando este último relacionado con los ecosistemas y vida submarina.

TABLA 4.24. CLASIFICACIONES ASBTRACTS ODS14 - MODELO

23

Objetivo	Clasificaciones
ODS1	1
ODS2	35
ODS3	64
ODS4	0
ODS5	0
ODS6	589
ODS7	5
ODS8	0
ODS9	9
ODS10	1
ODS11	43
ODS12	58
ODS13	97
ODS14	118
ODS15	65
ODS16	1
ODS17	4

Análisis objetivo 15

Este objetivo consigue un número más alto de clasificaciones, 411, cuadro 4.18. Esta, aunque alta, es más baja de lo que sería ideal. Esto como se aprecia en la figura 4.26, puede ser debido a la división de la taxonomía en dos grandes grupos. Además esto concuerda con la cantidad de textos asignados, ya que estos son aproximadamente la mitad del total. Unos de los dos grupos, el verde, habla de especies y cambio climático, temas que se podrían considerar como relacionados al objetivo 15. Adicionalmente el otro grupo, el rojo, contiene términos como sistema complejo y adaptativo y sostenibilidad, los cuales se pueden considerar más genéricos.

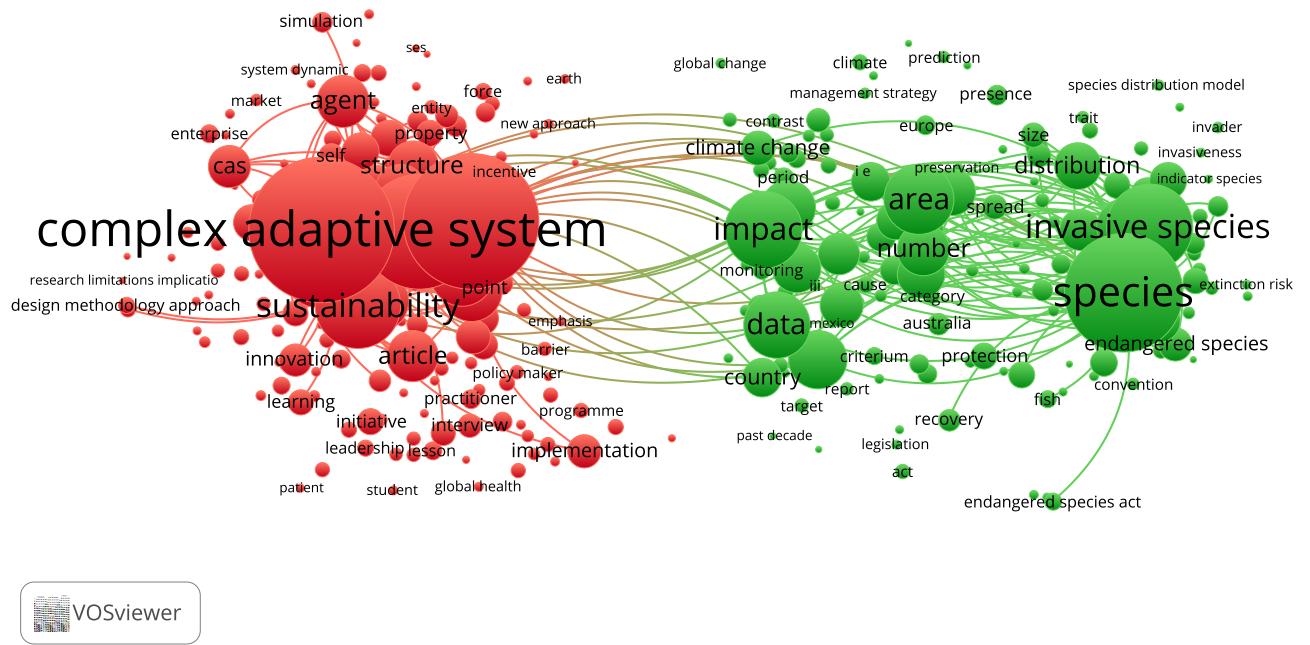


Fig. 4.26. Mapa de relaciones del objetivo 15

Análisis objetivo 16

Este objetivo obtiene, al igual que el 15, unos resultados positivos, más aún si se tiene en cuenta que el número total de textos es menor, pero el número de clasificaciones es similar, cuadro 4.18. Esto está reforzado por la figura 4.27, en la cual se aprecia una taxonomía altamente relacionada con el objetivo 16, cubriendo temas como los conflictos armados, guerras civiles, entre otros. Esto cuadra correctamente con los resultados obtenidos.

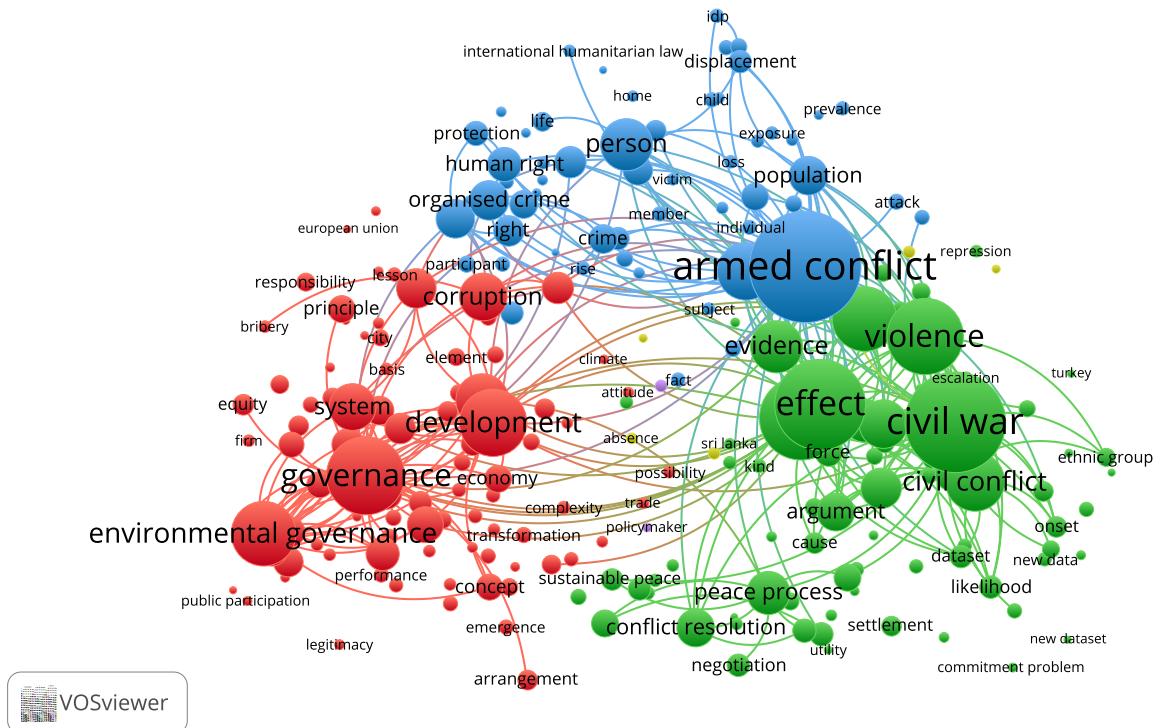


Fig. 4.27. Mapa de relaciones del objetivo 16

Análisis objetivo 17

Este último objetivo, he obtenido unas métricas razonables pero aún así relativamente bajas, cuadro 4.18. Esto se ve reflejado en la figura 4.28, presentando este una taxonomía altamente diversificada, tocando temas de todos los tipos como lo son la industria, el comercio, la pobreza y de todos ellos, el único realmente relacionado con el objetivo 17 es la colaboración, presente en un segundo plano en el grupo rojo de la figura.

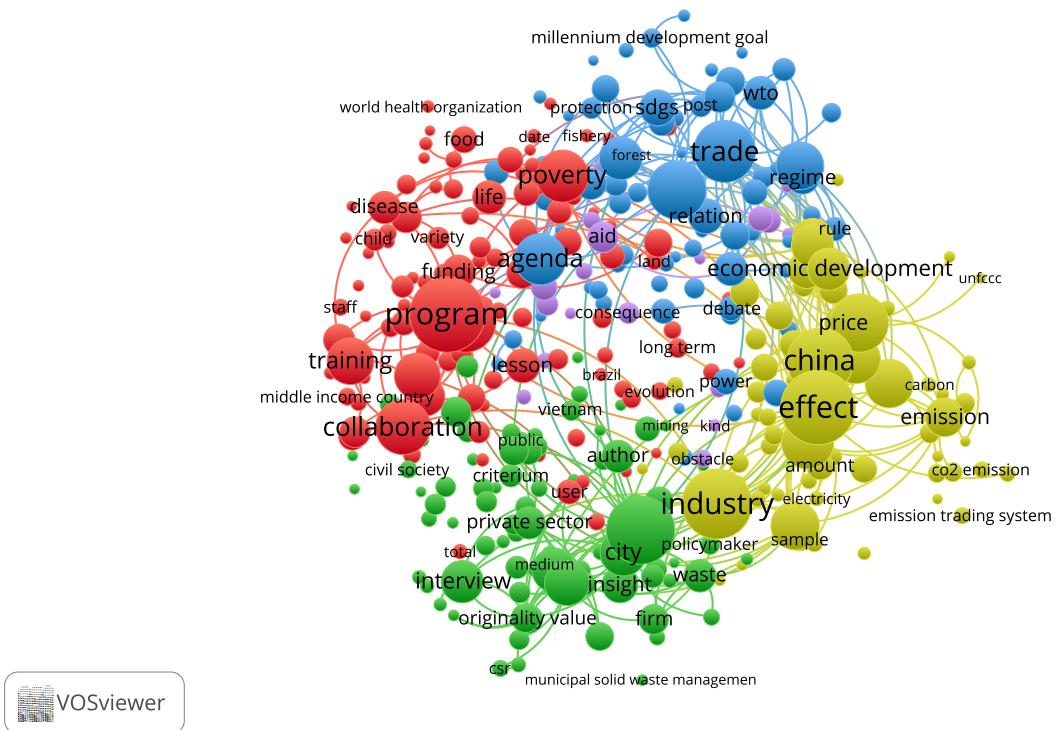


Fig. 4.28. Mapa de relaciones del objetivo 17

Para confirmar la diversidad en la taxonomía de los textos de este objetivo, se han incluido, en la tabla, cuadro 4.25, las clasificaciones del modelo 23 de todos los textos pertenecientes al conjunto de datos de este objetivo. En esta se ve como hay multitud de objetivos con en torno a 100-150 textos asignados y de todos el que más tiene es el 17 contando con cerca de 300. Esto concuerda correctamente con lo visto en la taxonomía, figura 4.28, siendo esta bastante variada.

TABLA 4.25. CLASIFICACIONES ASBTRACTS ODS17 - MODELO

23

Objetivo	Clasificaciones
ODS1	43
ODS2	95
ODS3	159
ODS4	21
ODS5	9
ODS6	47
ODS7	93
ODS8	68
ODS9	125
ODS10	16
ODS11	65
ODS12	109
ODS13	129
ODS14	23
ODS15	59
ODS16	5
ODS17	298

4.3.4. Conclusiones parciales

Gracias a estas pruebas podemos destacar el rendimiento del modelo 23, obteniendo este unas clasificaciones, en la mayoría de los casos correctas y en las que no, un posterior análisis de los datos usados demostró que realmente el modelo funcionaba según lo esperado pero los conjuntos no contenían datos relacionados con el objetivo esperado. Este último análisis también ha ayudado a determinar que resultados excepcionalmente buenos, como el del modelo 21, cuadro 4.17, realmente eran demasiado buenos como para ser verdad, analizando únicamente la tabla puede parecer que este modelo destaca por su rendimiento excepcional, obteniendo una métricas muy superiores al resto, pero si se analizan las taxonomías de los conjunto de datos , figuras 4.19 a 4.28, se observa que realmente un comportamiento correcto mostraría menos textos clasificados de manera correcta.

4.4. Conclusiones generales

Las pruebas realizadas pueden considerarse, en gran medida, como satisfactorias, aclarando luz sobre el funcionamiento y razonamiento detrás de cada modelo. Han ayudado a determinar si las métricas obtenidas por los modelos son correctas o no y finalmente ayudando a determinar cual de todos los modelos presenta un comportamiento superior, como es el caso del modelo 23.

Se puede argumentar que tal nivel de pruebas y de análisis puede llegar a ser excesivo pero si el objetivo finales realizar un análisis cuantitativo, como es el caso, la validez de los datos finales presentados depende directamente de la validación realizada y, como es obvio de los resultados de la misma, y de que esta sea suficientemente extensa y detallada.

5. RESULTADOS

En este apartado se presentan los resultados obtenidos en la realización de un estudio cuantitativo sobre la literatura científica de los últimos años. Dichos resultados no se analizarán en profundidad, conteniendo este apartado únicamente, una descripción de los mismos.

En este estudio se han clasificado multitud de textos referentes a la sostenibilidad, para analizar si existe algún patrón u objetivo predominante en este ámbito de estudio.

Como primer paso se han extraído un total de 100.000 textos de *Scopus*, una de las principales bases de datos académicas. Estos textos están divididos de manera equitativa entre los años 2019 y 2023, contando con 20.000 de cada año. Este número es debido a que esta es la cantidad máxima que la plataforma permite exportar. Seguramente esta limitación pueda ser abordada de alguna manera, pudiendo obtener todas las publicaciones de cada año. Esto se consideró en un principio, pero más tarde se llegó a la conclusión de que 20.000 era una población suficiente de datos como para poder identificar, si existe, algún patrón en las publicaciones. Adicionalmente el rango de años elegido ha sido de 5 años ya que se considera que este es suficiente para poder, por un lado, obtener una métricas relevantes en cuanto al esfuerzo investigador realizado en cada objetivo y por otro poder identificar alguna tendencia en los datos, teniendo en cuenta el impacto que la pandemia de COVID19 ha tenido en la sociedad.

Estos textos fueron clasificados usando el modelo que mejor resultados obtuvo en las validaciones, siendo este el 23. Los resultados obtenidos se encuentran a continuación representados en la tabla cuadro 5.1

TABLA 5.1. DATOS RESULTANTES DEL ESTUDIO

ODS/Año	2019	2020	2021	2022	2023	Total
ODS1	163	160	180	155	170	828
ODS2	2063	2273	2498	2354	2304	11492
ODS3	1742	1859	2002	2057	1884	9544
ODS4	511	535	459	472	386	2363
ODS5	256	255	261	243	221	1236
ODS6	1012	1068	1062	980	984	5106
ODS7	1777	1730	1885	1900	2021	9313
ODS8	908	874	880	927	884	4473
ODS9	2031	1960	1964	2219	2447	10621
ODS10	109	103	83	91	94	480
ODS11	2099	1922	1776	1684	1616	9097
ODS12	4405	4550	4605	5012	5361	23933
ODS13	1304	1311	1486	1633	1767	7501
ODS14	801	816	887	755	738	3997
ODS15	2337	2300	2303	2198	2078	11216
ODS16	161	166	145	115	128	715
ODS17	1829	1765	1641	1702	1688	8625

Esta tabla, cuadro 5.1, contiene los datos de manera condensada pero su interpretación puede ser complicada, es por esto por lo que se presentarán diversas representaciones de los datos en forma de gráficas para su fácil interpretación.

5.1. Análisis de resultados generales

Como primera representación de los resultados se incluye la gráfica, figura 5.1, la cual contiene el número total de textos clasificados por objetivo, entre los años 2019 y 2023. En esta se puede apreciar de manera más notable que el objetivo con más textos relacionados es el 12, seguido de una serie de objetivos con una relevancia similar, los 2, 3, 7, 9, 11, 13, 15 y 17. Estos tienen un número de publicaciones similares. Destacan también los objetivos 1, 10 y 16 por el número tan bajo de publicaciones relacionadas.

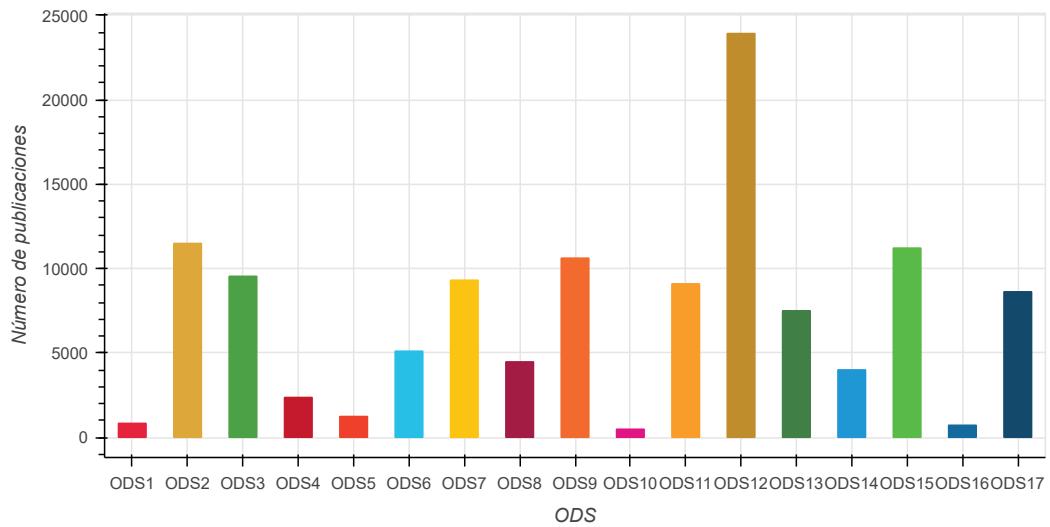


Fig. 5.1. Resultados generales

5.2. Evolución de publicaciones

A continuación se muestra la evolución en el número de publicaciones de los objetivos más relevantes a lo largo de los años. La evolución del resto se encontrará al final del apartado de una manera más compacta, de esta manera se pueden apreciar las tendencias de estos objetivos.

Objetivo 12

Este primer objetivo es el más relevante en cuanto al número de publicaciones, duplicando en magnitud a los objetivos siguientes. En la figura 5.2 se puede apreciar la tendencia en el número de publicaciones de este objetivo y como esta va en aumento cada año, haciendo esto de una manera lineal y más o menos constante ya que cada año tiene más investigaciones que el anterior.

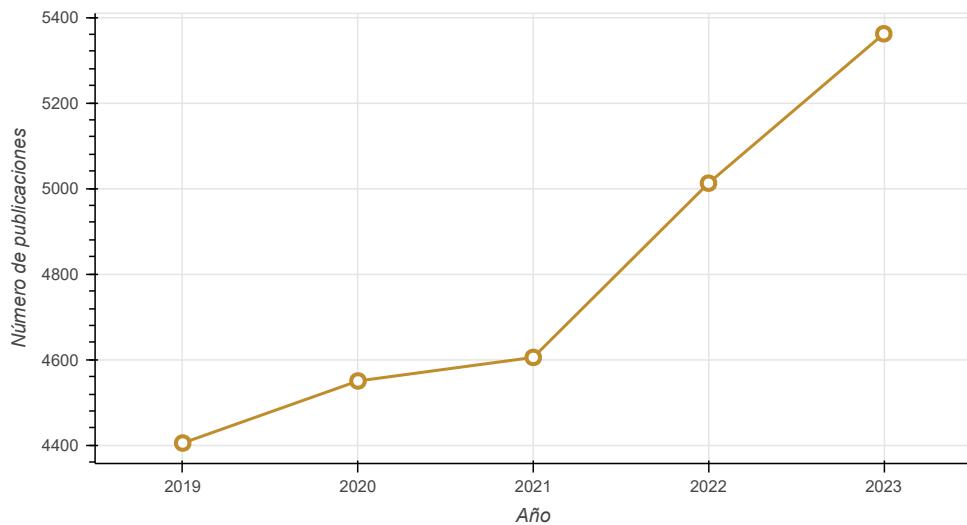


Fig. 5.2. Evolución ODS 12

Objetivo 3

Aunque no tan relevante en cuanto a métricas, el objetivo 3 tiene una relevancia especial debido a la pandemia generada por el coronavirus. El impacto que esta tuvo en la sociedad debería verse reflejado en un aumento en el número de publicaciones relacionadas con el objetivo 3. Esta correlación se ve reflejada en la figura 5.3, teniendo esta una tendencia ascendente y más acelerada que el objetivo 12 durante los años posteriores a 2019 pero esta se ve frenada de manera contundente en 2023. Aunque exista esta correlación solo con estos datos no se puede llegar a conclusiones en cuanto a causalidad pero la posibilidad tampoco se puede negar.

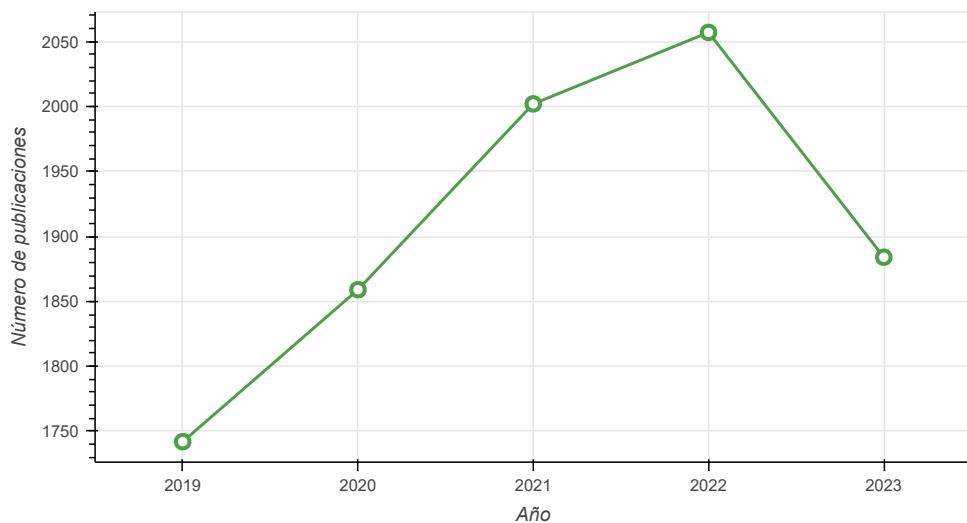


Fig. 5.3. Evolución ODS 3

Objetivo 9

La relevancia de este objetivo es constante, ya que la industria es uno de los principales motores económicos y de desarrollo humano, estando en los últimos años en la vanguardia tecnológica con el auge de la industria 4.0. Esta tendencia se ve reflejada en la gráfica, figura 5.4, estando el número de publicaciones relativamente estancado, aunque con un número alto de ellas, durante los años entre 2019 y 2021 y sufriendo un auge repentino a partir de 2022, manteniéndose esta tendencia a lo largo de 2023.

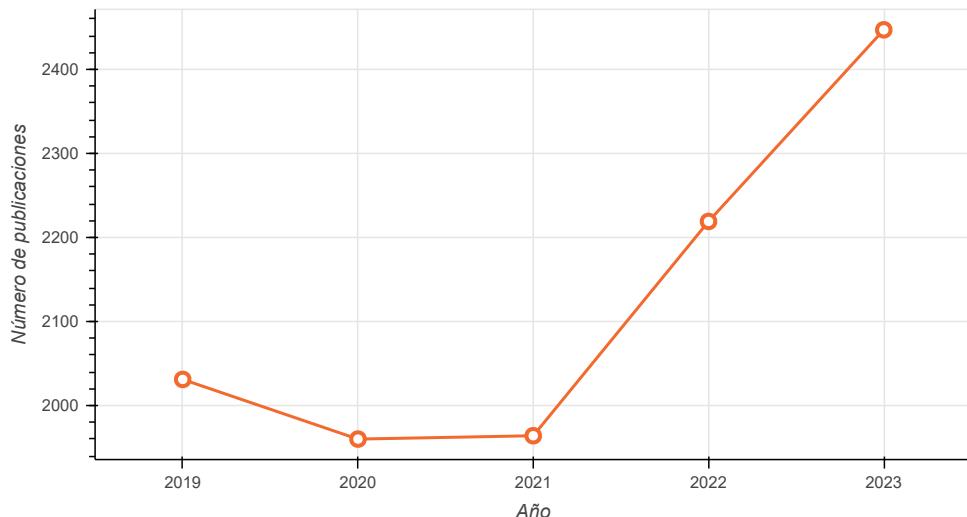


Fig. 5.4. Evolución ODS 9

Objetivo 2

La importancia de este objetivo en el contexto estudiado recae en su segundo lugar como objetivo con mayor número de publicaciones relacionadas, siendo estas, como se aprecia con mayor claridad en la tabla cuadro 5.1, de 11.492. Aún siendo la mitad que los obtenidos por el objetivo 12, es un número cuanto menos relevante. En la figura 5.5 se puede observar una tendencia ascendente desde 2019, con un pico en 2021 y una posterior tendencia descendiente hasta 2023.

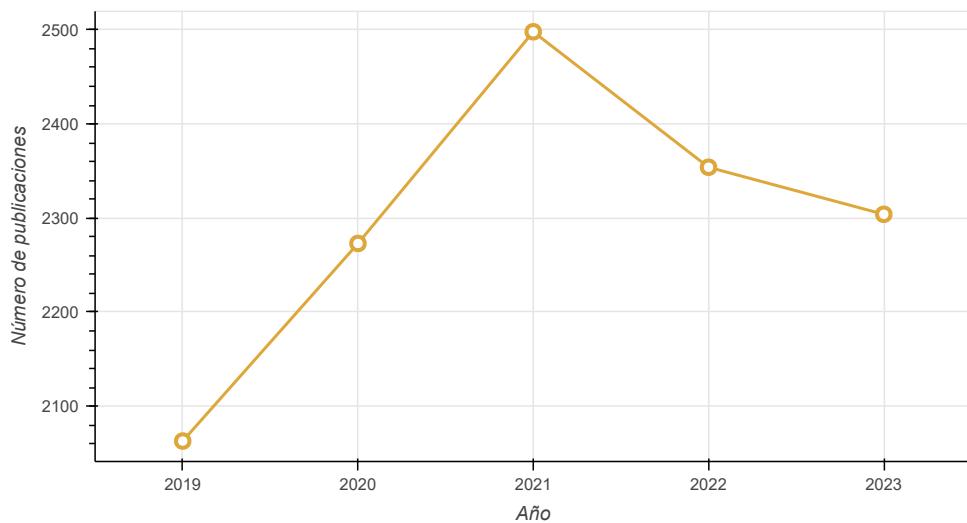


Fig. 5.5. Evolución ODS 2

Objetivo 15

Finalmente se incluye el objetivo 15 como último objetivo relevante, siendo este el tercero con más publicaciones relacionadas, por detrás del 2. Este, de todos los objetivos analizados, tiene la única gráfica que muestra una tendencia general descendiente, figura 5.6, con la notable excepción de las publicaciones de 2021, siendo estas ligeramente superiores a las de 2020. Resalta sobre todo la tendencia descendente tan drástica vista justo después de este aumento en publicaciones.

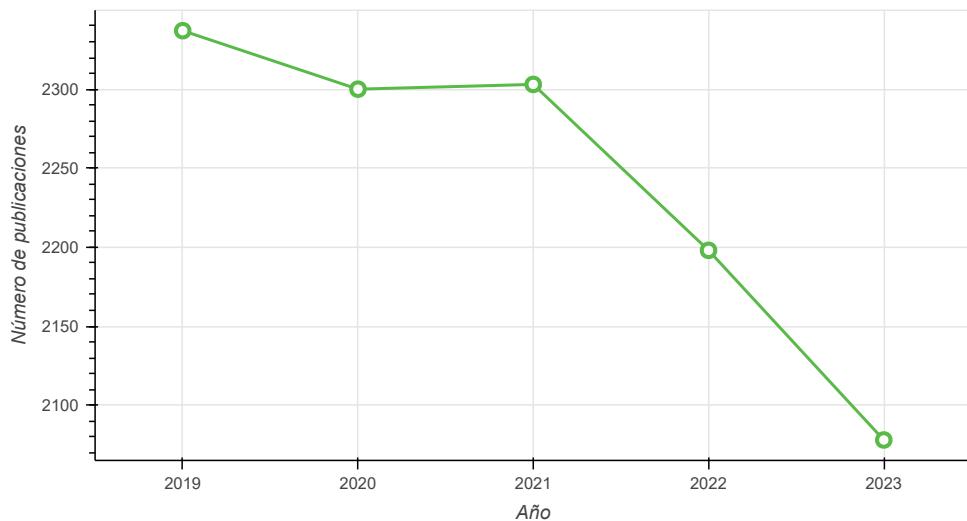
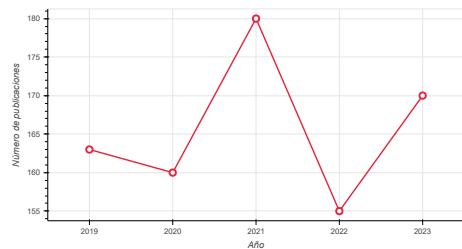


Fig. 5.6. Evolución ODS 15

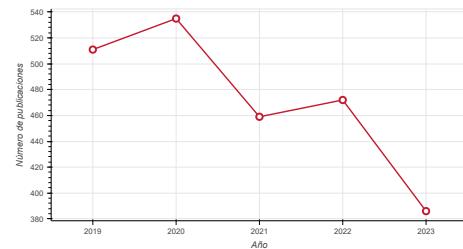
Objetivos adicionales

A continuación se muestran las gráficas de tendencias del resto de objetivos, figura 5.7, considerados menos relevantes como para mencionar y, por motivos de espacio y claridad,

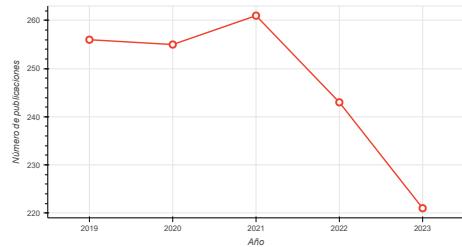
se incluirán en dos columnas. De esta manera las tendencias serán fácilmente distinguibles, pudiendo consultar los número en la tabla cuadro 5.1. Dentro de estas se ven multitud de tendencias diferentes, no pudiendo identificar ningún patrón común entre ellas. Destacar adicionalmente la gráfica referente al objetivo 11, figura 5.7h, siendo este uno de los objetivos con mayor número de publicaciones y contando con una tendencia descendente desde 2019.



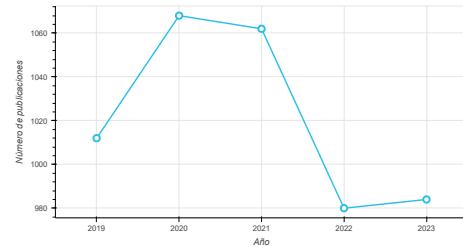
(a) Evolucion ODS1



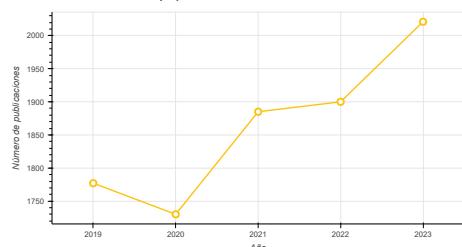
(b) Evolucion ODS4



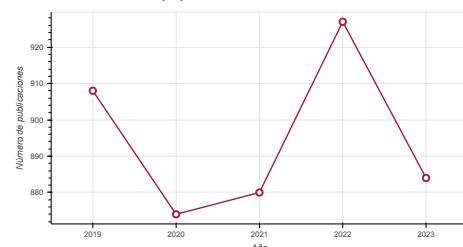
(c) Evolucion ODS5



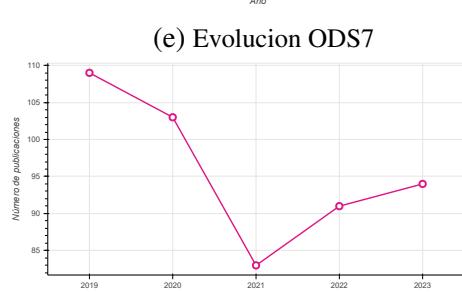
(d) Evolucion ODS6



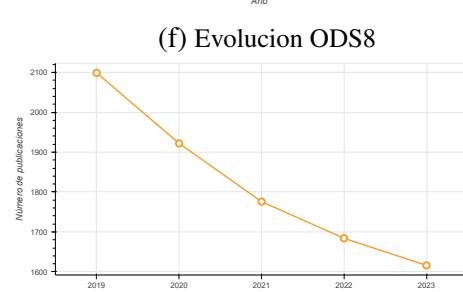
(e) Evolucion ODS7



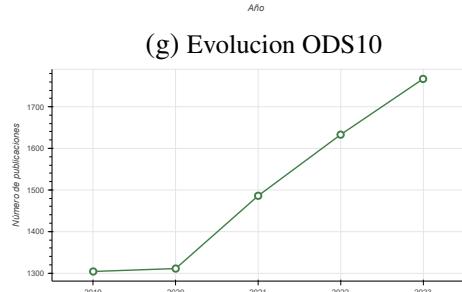
(f) Evolucion ODS8



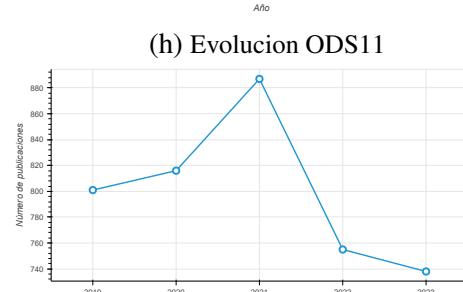
(g) Evolucion ODS10



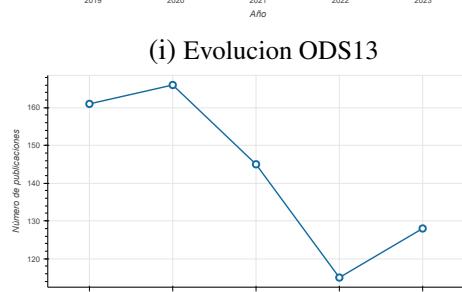
(h) Evolucion ODS11



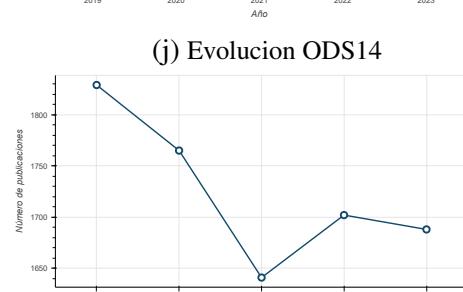
(i) Evolucion ODS13



(j) Evolucion ODS14



(k) Evolucion ODS16



(l) Evolucion ODS17

Fig. 5.7. Evolución del resto de objetivos

5.3. Proporciones generales

Como última representación de los datos se incluye, en la figura 5.8, un gráfico en forma de anillo en el que se muestran la proporción del número de clasificaciones de todos los objetivos en el conjunto general de datos. Este gráfico, representado en un formato inspirado en el logo original de los ODS, muestra las verdaderas relaciones de estos, dentro del contexto científico.

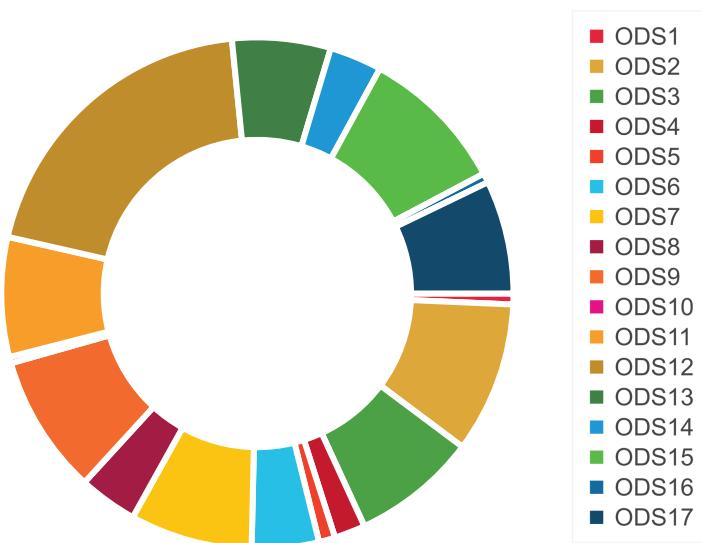


Fig. 5.8. Porcentaje de artículos por ODSa

6. GESTIÓN DEL PROYECTO

6.1. Planificación

Esta sección contiene las diferentes etapas del desarrollo, indicando de manera detallada el periodo que ocupa cada una en el plano general, las horas invertidas en cada una y los objetivos alcanzados.

Debido a la naturaleza secuencial del desarrollo, se han dividido todas las etapas en diferentes fases, en primer lugar se encuentran las fases previas al desarrollo, investigando el contexto del proyecto y el estado de la cuestión. En segundo lugar se encuentran las diferentes fases del desarrollo, algunas de las cuales precedidas por una pequeña fase de investigación en la que se investigaban los conceptos relacionados con las tecnologías a implementar en la fase que la sigue. En último lugar se encuentra una fase de validación general y el estudio final:

- **Análisis de contexto y proyecto:** Durante esta primera fase se exploró todo el entorno relacionado con el proyecto, realizando un estudio sobre los objetivos de desarrollo sostenible, explorando todo el contexto al rededor de los mismos y su importancia en la sociedad moderna.
- **Investigación del estado del arte:** Esta segunda fase se centró en el estudio del estado del arte, analizando las múltiples soluciones similares existentes y sus características principales. Adicionalmente se realizó un estudio sobre las diferentes arquitecturas y teoría relacionada con las redes de neuronas
- **Iteración inicial:** Esta primera fase de desarrollo está formada por tres etapas diferentes la primera de ellas consiste en una búsqueda de datos que usar para el entrenamiento. Esta es continuada por la generación y entrenamiento de una serie de modelos que finalmente fueron sometidos a una fase de validación y pruebas para analizar el progreso conseguido.
- **Investigación segunda iteración:** en la segunda iteración se decidió implementar una arquitectura nueva, basada en transformes, dándole un carácter vanguardista al proyecto, implementando las últimas tecnologías desarrolladas. Es por esto por lo que se investigó sobre el tema relacionado con los transformes, y más concretamente los modelos de aprendizaje por transferencia como BERT.
- **Segunda iteración:** Tras la investigación previa necesaria se procedió a entrenar una serie de estos nuevos modelos, los cuales fueron expuestos a la misma fase final de validación y pruebas.

- **Investigación tercera iteración:** Tras entrenar modelos con esta nueva arquitectura y analizar su rendimiento se decidió adaptar la base de datos para hacerla más adecuada a la hora de resolver tareas de clasificación multi-etiqueta. Es por esto por lo que en esta fase se estudiaron las diferentes técnicas de aumentado de datos textuales.
- **Tercera iteración:** Con los conocimientos adquiridos en la fase anterior, se procedió a aumentar la base de datos, generando una nueva mayor en magnitud y diseñada con la resolución de problemas multi-etiqueta en mente. Con esta nueva base de datos se repitieron los procesos de entrenamiento y validación de las fases anteriores.
- **Cuarta iteración:** Esta cuarta iteración surgió del descubrimiento de nuevos datos multi-etiquetados de manera natural, sustituyendo de esta forma la base de datos generada de manera artificial. Al igual que en las fases anteriores, con el nuevo conjunto de datos se entrenaron y validaron una serie de modelos finales.
- **Validación y pruebas final:** Como fase final de desarrollo se realizaron una serie de validaciones y pruebas estandarizadas sobre todos los modelos generados, consiguiendo así unas métricas generales que poder usar a modo de comparación entre modelos de fases diferentes.
- **Estudio cuantitativo:** Finalmente se realizó un estudio cuantitativo. Como primera etapa de este estudio se recogieron multitud de abstracts almacenados en bases de datos académicas. A continuación se clasificaron haciendo uso del modelo que presentó los mejores resultados en la fase de validación final y finalmente se interpretaron los resultados de dicho estudio.

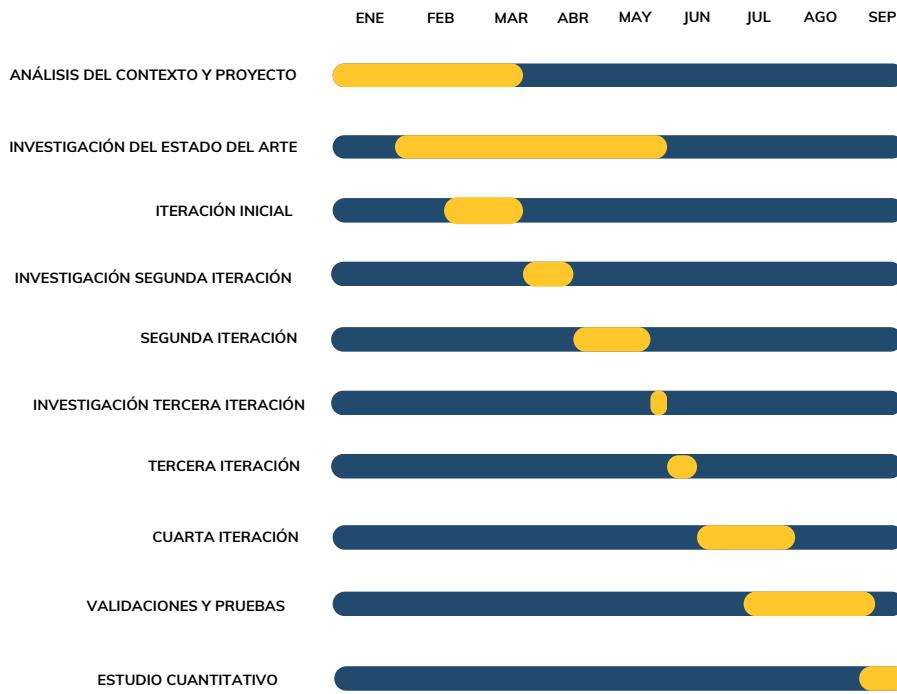


Fig. 6.1. Diagrama de Gantt

6.1.1. Presupuesto:

Esta sección contiene el desglose de los diversos recursos, humanos y materiales, usados durante el desarrollo del proyecto, incluyendo al final la suma de los costes totales.

Costes de Personal

Los primeros costes a tener en cuenta son los del personal, en el se incluyen las horas de trabajo del alumno junto con las de tutorías y de trabajo por parte del tutor, que en este caso toma el papel de director del proyecto. El total de horas que corresponden al trabajo por parte del alumno son 375, originadas del número de créditos asignados al proyecto, siendo estos un total de 12 y equivaliendo cada uno a 31,25 horas de trabajo se llega a esta cifra final. El coste total de personal se obtiene multiplicando el número de horas finales por los sueldos estándares de cada rol, siendo estos 15€/hora en el caso de un ingeniero y de 50€/h en el caso del director.

TABLA 6.1. TABLA COSTES DE PERSONAL.

Cargo	Horas	Coste por hora	Coste Total
Ingeniero	375	15€/hora	5.625€
Director del Proyecto	25	50€/hora	750€
Total			6.375€

Costes Software y Hardware

Los costes relacionados con el software y hardware utilizados son los siguientes, teniendo en cuenta la compra de un portátil (HP Pavilion Notebook 15-bc5) al inicio de la carrera, el gasto parcial de este será el correspondiente a los 4 meses de trabajo dentro de los 4 años totales de uso. El resto de gastos de hardware y software son 0 debido al uso de herramientas y software *open-source* y a los beneficios obtenidos por ser estudiante, los cuales proporcionan multitud de licencias gratis.

TABLA 6.2. TABLA COSTES DE SOFTWARE Y HARDWARE.

Producto	Coste	Amortización	Coste por mes	Uso en proyecto	Coste Total
Portátil	500€	48 meses	10,41 $\frac{\text{€}}{\text{mes}}$	4 meses	41,64€
VisualStudio Code	0	-	-	-	0
Python	0	-	-	-	0
Windows	0	-	-	-	0
Total	-	-	-	-	41,64€

Costes Indirectos

Los costes indirectos son aquellos que corresponden a los recursos no materiales utilizados durante el proyecto, estos incluyen la electricidad e internet usados.

TABLA 6.3. TABLA COSTES INDIRECTOS.

Producto	Precio por Mes	Uso en Proyecto	Coste Total
Electricidad	50€	6 meses	300€
Servicio de Internet	30€	6 meses	180€
Total			480€

Finalmente se incluyen los costes totales, donde se agrupan todos los anteriores para generar un presupuesto final del proyecto.

TABLA 6.4. TABLA COSTES TOTALES.

Producto	Gasto Total
Personal	6.375€
Material	41,64€
Costes indirectos	480€
Total	6.896,64€

6.2. Impacto socio-económico

El desarrollo de este proyecto tiene un potencial impacto socio-económico considerable. Este impacto se extiende a varias áreas críticas y se encuentra en consonancia con los

esfuerzos globales hacia la sostenibilidad, tal como lo establece la Agenda 2030 de las Naciones Unidas. En este contexto, la asignación automatizada de ODS a textos puede ser vista como un valioso recurso para mejorar la eficiencia en la gestión y la toma de decisiones, así como para promover la conciencia y la acción en torno a los ODS.

En primer lugar, este tipo de proyecto respalda directamente la Agenda de Desarrollo Sostenible de las Naciones Unidas, que establece 17 ODS para abordar problemas globales como la pobreza, el hambre, la igualdad de género, la acción climática y la paz y la justicia. Al asignar los ODS de manera automatizada a documentos y textos relacionados, el proyecto facilita la alineación de esfuerzos y recursos hacia la consecución de estos objetivos, contribuyendo así al progreso global hacia un futuro sostenible.

Además, la asignación automatizada de ODS a textos puede tener un impacto significativo en la toma de decisiones informadas. Al identificar de manera eficiente y precisa la relevancia de los contenidos de texto en relación con los ODS, las organizaciones y los responsables de la toma de decisiones pueden contar con una herramienta valiosa para evaluar la coherencia y la contribución de sus acciones y políticas a los objetivos de sostenibilidad. Esta mejora en la capacidad de toma de decisiones puede tener repercusiones positivas tanto en el sector público como en el privado.

En el contexto de la investigación científica, la aplicación de sistemas de clasificación automatizada de textos emerge como una herramienta de gran relevancia para agilizar y optimizar la revisión de la literatura relacionada con los Objetivos de Desarrollo Sostenible (ODS). Esta tecnología posee el potencial de reducir de manera significativa el tiempo invertido por los investigadores en la búsqueda, selección y análisis de documentos relacionados con los ODS, lo que a su vez conduce a una mejora palpable en la calidad y eficiencia de la investigación en los campos asociados a la sostenibilidad.

La capacidad de estos sistemas de identificar y categorizar de manera rápida y precisa los documentos que abordan aspectos relevantes de los ODS, proporciona una ventaja sustancial en la identificación de áreas específicas que requieren una mayor atención y dedicación de recursos. Esto permite una asignación más efectiva de recursos humanos y financieros, facilitando la priorización de investigaciones en los ámbitos que requieren una mayor atención en términos de estudios o que presentan un mayor potencial de impacto en la consecución de los ODS.

Desde una perspectiva más amplia, la aplicación de modelos de asignación automatizada de Objetivos de Desarrollo Sostenible ODS a textos no solo influye en la toma de decisiones políticas y estratégicas al proporcionar datos cuantitativos para identificar áreas de investigación prioritarias, sino que también puede tener un impacto significativo en la conciencia pública y la adopción de prácticas sostenibles. Al etiquetar y resaltar contenido relacionado con los ODS, se puede fomentar una mayor comprensión y compromiso con los problemas y desafíos que abordan estos objetivos. Esto, a su vez, puede influir en las decisiones individuales y colectivas, así como en la forma en que las instituciones, públicas y privadas, abordan las cuestiones de sostenibilidad en sus operaciones

económicas y políticas.

Desde una perspectiva económica, este proyecto, se anticipa que tenga un impacto económico mínimo, cercano a cero, ya que no se enfoca en objetivos comerciales explícitos como la monetización del modelo. Su principal valor radica en su utilidad para la investigación y el análisis de los Objetivos de Desarrollo Sostenible (ODS) al automatizar el procesamiento y la clasificación de textos. Es este último aspecto el que puede tener cierto impacto económico, aunque limitado, pudiendo llegar a reducir el tiempo necesario para tratar textos, ayudando a alocar recursos económicos en otras áreas de interés.

En resumen, este proyecto puede tener un impacto socio-económico positivo y significativo. Esto se manifiesta en la promoción de la sostenibilidad, la mejora de la toma de decisiones, la eficiencia en la investigación científica, la generación de conocimiento y la promoción de la responsabilidad gubernamental y empresarial. Además, este proyecto se alinea con los objetivos globales de desarrollo sostenible y contribuye al avance de la Agenda 2030 de las Naciones Unidas.

7. CONCLUSIONES

En este apartado se presentan las conclusiones generales del trabajo, donde se recopilan todas lecciones aprendidas, al igual que se analizan los objetivos cumplidos. Adicionalmente se exponen las dificultades encontradas en el desarrollo y las posibles líneas de trabajo futuro.

7.1. Conclusiones generales

Como se estableció en la introducción, el objetivo principal de este trabajo era el poder clasificar de manera efectiva textos acorde a los ODS relacionados con el mismo. Después de un largo desarrollo y una exhaustiva investigación se puede afirmar que este ha sido cumplido.

El desarrollo de una serie de modelos de aprendizaje profundo resulta un aprendizaje importante, más aún si se tiene en cuenta que se han seguido diversas arquitecturas y estructuras de datos por lo que, de manera adicional, se ha ayudado a entender mejor el impacto y efecto que estos aspectos tienen en el rendimiento final del modelo. Parte esencial esta evaluación es la exhaustiva validación que se ha llevado a cabo, parte esencial de la cual ha sido el uso de herramientas de análisis de datos que ayuden a entender de manera más profunda la estructura de estos demostrando al importancia de realizar un análisis de este tipo antes de analizar de manera final los resultados y no saltar a conclusiones de manera acelerada.

El recopilado de datos de internet ha demostrados ser una herramienta extremadamente útil, abriendo un mundo de posibilidades infinito, proporcionando una cantidad inmensa de datos. Destacar también la importancia de realizar una extracción legítima de estos datos, siempre en concordancia con lo que dictamina el marco regulador.

La investigación llevada a cabo al final, aunque menor en magnitud, representa un hito importante, demostrando la capacidad e importancia del modelos como los desarrollados en este proyecto a la hora de tratar con cantidades grandes y heterogéneas de datos. Y demostrando la posibilidad de adaptar estos modelos a temas con taxonomías extensas y complicadas y obteniendo aún así, unos resultados que, teniendo en cuenta los resultados de las validaciones, resultan fiables.

7.2. Dificultades y limitaciones

El desarrollo de modelos de aprendizaje profundo está plagado de limitaciones, en primer lugar se encuentran las limitaciones de hardware ya que, el uso de modelos grandes resulta una tarea pesada que no puede ser llevada a cabo por cualquier ordenador, el

adaptar los modelos a estas limitaciones es siempre una tarea difícil y más aún cuando no se cuenta con un hardware puntero. En segundo lugar se encuentran las limitaciones de software ya que los entornos de desarrollo utilizados son altamente dependientes en la versión usada, si se implementa alguna característica relativamente novedosa esto podría dar lugar a incompatibilidades entre funcionalidades y módulos diferentes.

Otra limitación referente a la creación de modelos de aprendizaje profundo, y aplicable a cualquier tarea de aprendizaje automático, es la recolección de los datos, esto puede ser una tarea difícil si no se cuenta con una base de datos extensa inicial. El recopilado de datos es una buena solución pero tiene que llevarse a cabo con cautela debido a la alta relación entre los datos usados y los resultados obtenidos.

7.3. Líneas de trabajo futuro

Este trabajo se deja diversas funcionalidades, características e investigaciones en el tintero, abriendo un amplio abanico de líneas futuras de trabajo. Como línea principal se encuentra el adaptar la clasificación de únicamente textos en inglés a textos en otros idiomas como puede ser el español, facilitando así el monitoreo de diferentes entornos e instituciones.

De manera adicional se podría hacer uso de modelos como los desarrollados en este trabajo para ampliar el estudio de las relaciones entre los ODS, pudiendo por ejemplo identificar la frecuencia en la que los diferentes objetivos aparecen juntos, identificando, de esta manera, como se relacionan los objetivos entre y las influencias que unos tienen sobre otros.

Otros ámbitos en los que se podría desarrollar un estudio como el desarrollado en este podrían ser el conjunto de regulaciones ambientales y de sostenibilidad, en el caso de España estas regulaciones son impuestas a nivel europeo por lo que un estudio sobre estas podría resultar útil. Pudiendo así comparar los esfuerzos científicos con los legislativos, identificando así diferencias o coincidencias de intereses.

ACRÓNIMOS

ADAM Adaptive Moment Estimation. 34, *Glosario*: Adaptive Moment Estimation

API Interfaz de Programación de Aplicaciones. 22, *Glosario*: Application Programming Interface

ASCII American Standard Code for Information Interchange. *Glosario*: American Standard Code for Information Interchange

BERT Bidirectional Encoder Representations from Transformers. 14, 21, 23, 30, 31, 34, 39, 43, 63, 96, *Glosario*: Bidirectional Encoder Representations from Transformers

CBOW Continuous Bag of Words. 17, *Glosario*: Continuous Bag of Words

CSV Comma Separated Values. 30, 32, 37, *Glosario*: Comma Separated Values

EDA Easy Data Augmentation. 37, *Glosario*: Easy Data Augmentation

GloVe Global Vectors for Word Representation. 23, *Glosario*: Global Vectors for Word Representation

HTML HTML. 21, 26, 27, *Glosario*: Lenguaje de Marcado de Hipertexto

HTTP Protocolo de Transferencia de HiperTexto. 20, 26, *Glosario*: Protocolo de Transferencia de HiperTexto

IA Inteligencia Artificial. 6, 8–10, 15, *Glosario*: Inteligencia Artificial

IISD International Institute for Sustainable Development. 37–39, 49, *Glosario*: International Institute for Sustainable Development

JOSS Journal of Open Source Software. 22, *Glosario*: Journal of Open Source Software

LDA Latent Dirichlet Allocation. 23, *Glosario*: Latent Dirichlet Allocation

NLP Procesamiento del Lenguaje Natural. 14, 15, *Glosario*: Procesamiento del Lenguaje Natural

ODS Objetivos de Desarrollo Sostenible. 1, 2, 6, 7, 21–23, 29, 33, 39, 95, 100–103, *Glosario*: Objetivos de Desarrollo Sostenible

ONU Organización de las Naciones Unidas. 6, *Glosario*: Organización de las Naciones Unidas

PDF Portable Document Format. 22, *Glosario*: Portable Document Format

SG Continuous Skip-gram. 17, *Glosario*: Continuous Skip-gram

TF-IDF Term Frequency-Inverse Document Frequency. 23, *Glosario*:

UN DESA Departamento de Asuntos Económicos y Sociales de las Naciones Unidas.
22, *Glosario*: Departamento de Asuntos Económicos y Sociales de las Naciones
Unidas

UNDP Programa de las Naciones Unidas para el Desarrollo. 38, *Glosario*: United Na-
tions Development Programme

UNEP Programa de las Naciones Unidas para el Medio Ambiente. 21, *Glosario*: Progra-
ma de las Naciones Unidas para el Medio Ambiente

UPM Universidad Politécnica de Madrid. 23, *Glosario*: Universidad Politécnica de Ma-
drid

USE Universal Sentence Encoders. 23, *Glosario*: Universal Sentence Encoders

UTF-8 Unicode Transformation Format 8-bit. 27, *Glosario*: Unicode Transformation
Format 8-bit

WoS Web of Science. 32, 39, 62, 67, *Glosario*: Web of Science

GLOSARIO

Adaptive Moment Estimation ADAM es un algoritmo de optimización utilizado en el aprendizaje automático y el entrenamiento de redes neuronales artificiales. Combina conceptos de otros algoritmos de optimización, como el descenso de gradiente estocástico (SGD) y RMSprop, para ajustar dinámicamente la tasa de aprendizaje durante el proceso de entrenamiento. Es conocido por su eficacia en la convergencia rápida y la adaptación a diferentes tasas de aprendizaje para diferentes parámetros del modelo. Es ampliamente utilizado en la optimización de redes neuronales profundas y otros modelos de aprendizaje automático.. 34

American Standard Code for Information Interchange El American Standard Code for Information Interchange (ASCII) es un conjunto de caracteres y códigos de control utilizados para representar texto y controlar dispositivos en sistemas informáticos y de telecomunicaciones. ASCII utiliza números de 7 bits para representar caracteres alfabéticos, numéricos y especiales, lo que permite la comunicación de texto en una amplia variedad de dispositivos y plataformas informáticas. Es un estándar ampliamente utilizado en la codificación de texto y proporciona una forma común de representar caracteres en la mayoría de las computadoras y sistemas de comunicación.. 27

Application Programming Interface Una API, o Interfaz de Programación de Aplicaciones, es un conjunto de reglas y protocolos que permite que diferentes software se comuniquen entre sí. Proporciona un conjunto de funciones y procedimientos que los desarrolladores pueden utilizar para interactuar con una aplicación o sistema informático específico. Las APIs son fundamentales para la integración de servicios y la creación de aplicaciones de software que aprovechan las funcionalidades de otros programas o plataformas.. 22

Bidirectional Encoder Representations from Transformers BERT se refiere a Bidirectional Encoder Representations from Transformers. Es un modelo de lenguaje desarrollado por Google en 2018 para tareas de procesamiento del lenguaje natural (NLP, por sus siglas en inglés). Lo que distingue a BERT de otros modelos es su capacidad para entrenar representaciones de palabras basadas en su contexto en ambas direcciones (izquierda y derecha) de una palabra en una frase, lo que lo hace bidireccional. Esto permite a BERT tener un desempeño superior en muchas tareas de NLP, desde la comprensión de texto hasta la respuesta a preguntas.. 14

Continuous Bag of Words "CBOW" se refiere a "Continuous Bag of Words". Es un modelo utilizado en el entrenamiento de word embeddings, específicamente en el marco Word2Vec. CBOW predice palabras objetivo (por ejemplo, palabras centrales) a

partir de sus palabras de contexto circundantes en una ventana determinada, a diferencia del modelo Skip-Gram de Word2Vec que hace lo contrario: predice palabras de contexto a partir de una palabra objetivo. Ambos métodos son técnicas para generar representaciones vectoriales densas de palabras a partir de grandes corpus de texto.. 17

Continuous Skip-gram El modelo "Skip-gram", parte del marco Word2Vec, tiene como objetivo predecir palabras circundantes (contexto) a partir de una palabra dada. Es una técnica para generar representaciones vectoriales (embeddings) de palabras de manera que palabras con contextos similares queden cercanas en un espacio vectorial. . 17

Departamento de Asuntos Económicos y Sociales de las Naciones Unidas El Departamento de Asuntos Económicos y Sociales de las Naciones Unidas (UN DESA) es una entidad de las Naciones Unidas que se encarga de promover el desarrollo sostenible y la cooperación internacional en asuntos económicos y sociales. El Departamento desempeña un papel clave en la investigación y análisis de cuestiones económicas, sociales y ambientales a nivel global. También brinda apoyo técnico y asesoramiento a los estados miembros de las Naciones Unidas en áreas como el desarrollo económico, social y ambiental.. 22

Easy Data Augmentation Easy Data Augmentation (EDA) es una técnica utilizada en el procesamiento de lenguaje natural (NLP) y el aprendizaje automático para aumentar la cantidad de datos de entrenamiento. EDA consiste en aplicar transformaciones simples y controladas a los datos de entrenamiento existentes para generar nuevas instancias de datos. Estas transformaciones pueden incluir sinónimos, eliminación de palabras, inserción de palabras, y permutación de palabras, entre otros. EDA es útil para mejorar el rendimiento de los modelos de NLP al proporcionar más variabilidad en los datos de entrenamiento, lo que a menudo resulta en un mejor rendimiento en tareas como la clasificación de texto y la traducción automática.. 37

Global Vectors for Word Representation Global Vectors for Word Representation (GloVe) es un algoritmo de aprendizaje no supervisado que se utiliza para representar palabras en un espacio vectorial continuo. GloVe se utiliza comúnmente en el procesamiento del lenguaje natural (NLP) y la recuperación de información para capturar las relaciones semánticas entre palabras. Este algoritmo mapea palabras en vectores numéricos de manera que palabras con significados similares están ubicadas cerca unas de otras en el espacio vectorial. GloVe se utiliza ampliamente en tareas de NLP, como la traducción automática, el análisis de sentimientos y la recuperación de información.. 23

Inteligencia Artificial La IA, o Inteligencia Artificial, es un campo de la informática que se enfoca en desarrollar sistemas y programas que pueden realizar tareas que

normalmente requieren inteligencia humana, como el aprendizaje, la toma de decisiones, el reconocimiento de patrones y la resolución de problemas. Utiliza algoritmos y datos para simular procesos cognitivos y mejorar la automatización y el rendimiento en una amplia variedad de aplicaciones, desde asistentes virtuales hasta vehículos autónomos.. 6, 8, 20

International Institute for Sustainable Development El International Institute for Sustainable Development (IISD) es una organización sin fines de lucro dedicada a la promoción del desarrollo sostenible a nivel global. El IISD se enfoca en la investigación, análisis y promoción de políticas y prácticas que contribuyan a un desarrollo económico, social y ambiental equitativo y sostenible. La organización trabaja en una variedad de áreas, incluyendo el cambio climático, la conservación de recursos naturales, el comercio sostenible y la gobernanza ambiental. El IISD colabora con gobiernos, empresas y organizaciones de la sociedad civil para abordar los desafíos globales relacionados con la sostenibilidad.. 37

Journal of Open Source Software El "Journal of Open Source Software" (JOSS) es una revista académica con revisión por pares que se centra en la publicación de software de investigación de código abierto. Los artículos publicados en JOSS son breves y destinados principalmente a presentar y describir el software, su utilidad y relevancia en la investigación científica. La principal ventaja de publicar en JOSS es que promueve la práctica de hacer que el software científico sea de código abierto y revisable, lo que a su vez mejora la reproducibilidad y la transparencia en la investigación.. 22

Latent Dirichlet Allocation Latent Dirichlet Allocation (LDA) es un modelo generativo utilizado en el procesamiento de lenguaje natural y el análisis de texto. LDA es una técnica de aprendizaje no supervisado que se utiliza para descubrir temas latentes en un conjunto de documentos. El modelo asume que cada documento está compuesto por una mezcla de temas, y cada tema está compuesto por una mezcla de palabras. LDA se utiliza comúnmente en tareas como la agrupación de documentos, la clasificación de texto y la extracción de temas. Es una herramienta valiosa para analizar grandes colecciones de texto y descubrir patrones y estructuras subyacentes en los datos.. 23

Objetivos de Desarrollo Sostenible Los ODS, o Objetivos de Desarrollo Sostenible, son un conjunto de 17 metas globales establecidas por las Naciones Unidas en 2015 como parte de la Agenda 2030 para abordar problemas mundiales urgentes, como la pobreza, el hambre, la igualdad de género, la educación, la salud, el medio ambiente y la paz. Estos objetivos tienen como objetivo mejorar la calidad de vida de las personas y proteger el planeta, promoviendo un desarrollo sostenible a nivel económico, social y ambiental.. 1, 3, 6, 25, 100, 101

Organización de las Naciones Unidas La ONU, o Organización de las Naciones Unidas, es una organización internacional establecida en 1945 para promover la cooperación y la paz mundial. Su objetivo principal es facilitar la diplomacia y la resolución pacífica de conflictos entre países, así como abordar problemas globales como los derechos humanos, el desarrollo sostenible y la seguridad internacional. La ONU consta de 193 Estados miembros y desempeña un papel fundamental en la promoción de la cooperación internacional y la protección de los derechos y el bienestar de las personas en todo el mundo.. 6, 34

Portable Document Format El formato PDF, o Portable Document Format, es un formato de archivo diseñado para presentar documentos de manera independiente del software, el hardware y el sistema operativo utilizados para crearlos. Los archivos PDF pueden contener texto, imágenes, gráficos y otros elementos y mantienen su formato y diseño original cuando se visualizan en diferentes dispositivos y plataformas. Son ampliamente utilizados para compartir documentos, formularios, libros electrónicos y otros tipos de contenido digital.. 22

Programa de las Naciones Unidas para el Medio Ambiente La principal autoridad mundial en asuntos ambientales, establecida por las Naciones Unidas en 1972. Su objetivo es promover la conservación y el uso sostenible de los recursos naturales, así como abordar los desafíos ambientales globales. Trabaja en colaboración con gobiernos, organizaciones y la sociedad civil para proteger el medio ambiente y fomentar un desarrollo sostenible en todo el mundo.. 21

Term Frequency-Inverse Document Frequency TF-IDF, o Term Frequency-Inverse Document Frequency (Frecuencia de Término-Frecuencia Inversa de Documento), es una técnica utilizada en el procesamiento de texto y la recuperación de información. Se utiliza para evaluar la importancia de una palabra o término en un documento dentro de un corpus de documentos. TF-IDF combina dos métricas: la frecuencia de término (TF), que mide la frecuencia de una palabra en un documento específico, y la frecuencia inversa de documento (IDF), que mide la importancia de un término en todo el corpus. La puntuación TF-IDF se utiliza comúnmente para la representación de documentos y la recuperación de información, permitiendo identificar palabras clave o términos relevantes en un contexto dado.. 23

Unicode Transformation Format 8-bit UTF-8 es un estándar de codificación de caracteres que representa la mayoría de los caracteres del conjunto de caracteres Unicode utilizando secuencias de bytes de 8 bits. UTF-8 es un sistema de codificación ampliamente utilizado en la informática y la comunicación digital, ya que permite representar una amplia variedad de caracteres, incluyendo caracteres alfabéticos, numéricos, símbolos y caracteres especiales de diferentes idiomas y escrituras. Es compatible con ASCII y es el estándar de codificación predeterminado en muchas aplicaciones y sistemas operativos.. 27

United Nations Development Programme El Programa de las Naciones Unidas para el Desarrollo (PNUD o UNDP en inglés) es la red global de desarrollo de las Naciones Unidas. Ayuda a los países a alcanzar los Objetivos de Desarrollo Sostenible (ODS) y a eliminar la pobreza en todas sus formas y dimensiones. El PNUD trabaja en áreas como la reducción de la desigualdad, la promoción de la igualdad de género, la construcción de capacidades y la sostenibilidad ambiental.. 38

Universal Sentence Encoders Universal Sentence Encoders (USE) es un conjunto de modelos de procesamiento de lenguaje natural desarrollados por Google. Estos modelos están diseñados para convertir oraciones o fragmentos de texto en vectores numéricos, lo que facilita la comparación y el análisis de similitud semántica entre oraciones en tareas de procesamiento de lenguaje natural. Los USE se utilizan en una variedad de aplicaciones, como la búsqueda semántica, la recuperación de información y la clasificación de texto. Estos modelos capturan representaciones de alta calidad para oraciones en varios idiomas y se pueden utilizar en una amplia gama de aplicaciones de NLP.. 23

Universidad Politécnica de Madrid La Universidad Politécnica de Madrid (UPM) es una institución de educación superior ubicada en Madrid, España. Es una de las principales universidades técnicas de España y ofrece una amplia gama de programas académicos en áreas como la ingeniería, la arquitectura, la informática y las ciencias aplicadas. La UPM se dedica a la investigación y la formación en tecnología y ciencia, y desempeña un papel importante en la promoción de la innovación y el desarrollo tecnológico en el país.. 23

BIBLIOGRAFÍA

- [1] I. Aguilera, *TFG*, <https://github.com/IgnacioAguileraUC3M/TFG>, 2023.
- [2] Boletín Oficial del Estado. “Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.” (2018), [En línea]. Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>.
- [3] Boletín Oficial del Estado. “Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia.” (1996), [En línea]. Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-1996-8930>.
- [4] Parlamento Europeo y Consejo de la Unión Europea. “Directiva (UE) 2019/790 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital y por la que se modifican las Directivas 96/9/CE y 2001/29/CE (Texto pertinente a efectos del EEE).” (2019), [En línea]. Disponible en: <https://www.boe.es/doue/2019/130/L00092-00125.pdf>.
- [5] Parlamento Europeo y Consejo de la Unión Europea. “DIRECTIVA 2004/48/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 29 de abril de 2004 relativa al respeto de los derechos de propiedad intelectual (Texto pertinente a efectos del EEE).” (2004), [En línea]. Disponible en: <https://www.boe.es/doue/2004/195/L00016-00025.pdf>.
- [6] Parlamento Europeo y Consejo de la Unión Europea. “Directiva 2011/77/UE del Parlamento Europeo y del Consejo, de 27 de septiembre de 2011 , por la que se modifica la Directiva 2006/116/CE relativa al plazo de protección del derecho de autor y de determinados derechos afines.” (2011), [En línea]. Disponible en: <https://www.boe.es/doue/2011/265/L00001-00005.pdf>.
- [7] D. o. E. United Nations y S. A. S. Development. “Transforming our world: the 2030 Agenda for Sustainable Development.” (2015), [En línea]. Disponible en: <https://sdgs.un.org/2030agenda>.
- [8] “Joint SDG Fund,” Joint SDG Fund. (2023), [En línea]. Disponible en: <https://www.jointsdgfund.org/es>.
- [9] D. o. E. United Nations y S. A. S. Development. “UN Sustainable Development Goals.” (2020), [En línea]. Disponible en: <https://sdgs.un.org/>.
- [10] “Progress Reports on the Sustainable Development Goals,” United Nations. (2023), [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/progress-report/>.

- [11] “What Is AI?” McKinsey & Company. (abr. de 2023), [En línea]. Disponible en: <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai#/>.
- [12] P. Ongsulee, “Artificial intelligence, machine learning and deep learning,” en *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2017, pp. 1-6. doi: [10.1109/ICTKE.2017.8259629](https://doi.org/10.1109/ICTKE.2017.8259629).
- [13] “The History of Artificial Intelligence,” Harvard University. (ago. de 2017), [En línea]. Disponible en: <https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>.
- [14] D. CSV. “¿Qué es un TRANSFORMER? La Red Neuronal que lo cambió TODO!” (Sep. de 2021), [En línea]. Disponible en: <https://www.youtube.com/watch?v=aL-EmKuB078>.
- [15] A. Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [16] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. arXiv: [1810.04805](https://arxiv.org/abs/1810.04805) [cs.CL].
- [17] D. Jurafsky y J. H. Martin. “Speech and Language Processing (3rd edition).” (2023), [En línea]. Disponible en: <https://web.stanford.edu/~jurafsky/slp3/>.
- [18] A. C. Vásquez, J. P. Quispe, A. M. Huayna et al., “Procesamiento de lenguaje natural,” *Revista de investigación de Sistemas e Informática*, vol. 6, n.º 2, pp. 45-54, 2009.
- [19] B. Li, P. Lu y v chmcl. “Normalize Data - Azure Machine Learning.” (nov. de 2021), [En línea]. Disponible en: <https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/normalize-data?view=azureml-api-2>.
- [20] D. L. Yse. “Text Normalization for Natural Language Processing (NLP).” (feb. de 2021), [En línea]. Disponible en: <https://towardsdatascience.com/text-normalization-for-natural-language-processing-nlp-70a314bfa646>.
- [21] T. Mikolov, K. Chen, G. Corrado y J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [22] Tensorflow. “word2vec | Text | Tensorflow.” (jul. de 2023), [En línea]. Disponible en: <https://www.tensorflow.org/text/tutorials/word2vec>.
- [23] P. Blunsom, K. Kocić y J. R. Curran, “Question Classification with Log-Linear Models,” en *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ép. SIGIR ’06, Seattle, Washington, USA: Association for Computing Machinery, 2006, pp. 615-616. doi: [10.1145/1148170.1148282](https://doi.org/10.1145/1148170.1148282). [En línea]. Disponible en: <https://doi.org/10.1145/1148170.1148282>.

- [24] A. Maiza. “Multi-Label Image Classification in TensorFlow 2.0.” (dic. de 2019), [En línea]. Disponible en: <https://towardsdatascience.com/multi-label-image-classification-in-tensorflow-2-0-7d4cf8a4bc72>.
- [25] S. Sharma, S. Sharma y A. Athaiya, “Activation functions in neural networks,” *Towards Data Sci*, vol. 4, n.º 12, pp. 310-316, 2020.
- [26] V. Liermann, S. Li y N. Schaudinnus, “Deep Learning: An Introduction,” en *The Impact of Digital Transformation and FinTech on the Finance Professional*, V. Liermann y C. Stegmann, eds. Cham: Springer International Publishing, 2019, pp. 305-340. doi: [10.1007/978-3-030-23719-6_17](https://doi.org/10.1007/978-3-030-23719-6_17). [En línea]. Disponible en: https://doi.org/10.1007/978-3-030-23719-6_17.
- [27] S. Narayan, “The generalized sigmoid activation function: Competitive supervised learning,” *Information Sciences*, vol. 99, n.º 1, pp. 69-82, 1997. doi: [https://doi.org/10.1016/S0020-0255\(96\)00200-9](https://doi.org/10.1016/S0020-0255(96)00200-9). [En línea]. Disponible en: <https://www.sciencedirect.com/science/article/pii/S0020025596002009>.
- [28] B. Zhao, “Web scraping,” *Encyclopedia of big data*, vol. 1, 2017.
- [29] Emeritus. “Here’s Why to Use Python for Data Science.” (2023), [En línea]. Disponible en: <https://emeritus.org/in/learn/heres-why-use-python-for-data-science/>.
- [30] L. Eggleston. “How Is Python Used for Data Science?” (2021), [En línea]. Disponible en: <https://www.courserreport.com/blog/how-is-python-used-for-data-science-metis>.
- [31] J. E. Guisiano, R. Chiky y J. De Mello, “SDG-Meter: A Deep Learning Based Tool for Automatic Text Classification of the Sustainable Development Goals,” en *Intelligent Information and Database Systems*, N. T. Nguyen et al., eds., Cham: Springer International Publishing, 2022, pp. 259-271.
- [32] Y. Li et al., “SDGdetector: an R-based text mining tool for quantifying efforts toward Sustainable Development Goals,” *Journal of Open Source Software*, vol. 8, n.º 84, p. 5124, 2023. doi: [10.21105/joss.05124](https://doi.org/10.21105/joss.05124). [En línea]. Disponible en: <https://doi.org/10.21105/joss.05124>.
- [33] L. Pukelis, N. B. Puig, M. Skrynik y V. Stanciauskas, “OSDG—Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs),” *arXiv preprint arXiv:2005.14569*, 2020.
- [34] L. Pukelis et al., “OSDG 2.0: a multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs),” *arXiv preprint arXiv:2211.11252*, 2022.
- [35] O. S. D. S. Group. “Open Source Data Science Group.” (2022), [En línea]. Disponible en: <https://osdg.ai/>.
- [36] U. N. S. D. Group. “Linked Sustainable Development Goals (SDGs).” (n.d.), [En línea]. Disponible en: <https://linkedsdg.officialstatistics.org/#/>.

- [37] United Nations Sustainable Development Group. “About the Linked Sustainable Development Goals (SDGs).” (n.d.), [En línea]. Disponible en: <https://sustainabledevelopment.un.org/LinkedSDGs/about>.
- [38] S. Meier, R. Mata y D. U. Wulff, “Text2sdg: An open-source solution to monitoring sustainable development goals from text,” *arXiv preprint arXiv:2110.05856*, 2021.
- [39] D. S. Meier, R. Mata y D. U. Wulff, “text2sdg: An R package to Monitor Sustainable Development Goals from Text,” *arXiv preprint arXiv:2110.05856*, 2021.
- [40] D. S. Meier, R. Mata y D. U. Wulff. “Text2SDG: A Multilingual and Multimodal Approach to Improve SDG Monitoring.” (), [En línea]. Disponible en: <https://www.text2sdg.io/>.
- [41] J. Pérez Álvarez, “Clasificador de textos de acuerdo con los Objetivos de Desarrollo Sostenible usando Procesamiento de Lenguaje Natural,” No Publicado, Madrid, España, 2021. [En línea]. Disponible en: <https://oa.upm.es/68099/>.
- [42] T. Matsui et al., “A natural language processing model for supporting sustainable development goals: translating semantics, visualizing nexus, and connecting stakeholders,” *Sustainability Science*, vol. 17, n.º 3, pp. 969-985, 2022.
- [43] F. Sovrano, M. Palmirani y F. Vitali, “Deep Learning Based Multi-Label Text Classification of UNGA Resolutions,” *CoRR*, vol. abs/2004.03455, 2020. arXiv: [2004.03455](https://arxiv.org/abs/2004.03455). [En línea]. Disponible en: <https://arxiv.org/abs/2004.03455>.
- [44] I. I. for Sustainable Development (IISD). “SDG Portal.” (n.d.), [En línea]. Disponible en: <https://sdg.iisd.org/>.
- [45] “United Nations Development Programme.” (n.d.), [En línea]. Disponible en: <https://www.undp.org/>.
- [46] N. Martens et al., “Anticipated benefits and challenges of implementing group care in Suriname’s maternity and child care sector: a contextual analysis,” *BMC Pregnancy and Childbirth*, vol. 23, n.º 1, 2023, Cited by: 0; All Open Access, Gold Open Access. doi: [10.1186/s12884-023-05904-y](https://doi.org/10.1186/s12884-023-05904-y). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168388230&doi=10.1186%2fs12884-023-05904-y&partnerID=40&md5=44677dab64bb6534dff72f8fa4611265>.
- [47] J.-I. N. Nneka et al., *Deforestation activities in ezekoro forest: Implications for climate change risks in Anambra state, southeast Nigeria*. 2024, 47 – 75, Cited by: 0. doi: [10.1002/9781119910527.ch3](https://doi.org/10.1002/9781119910527.ch3). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85165064040&doi=10.1002%2f9781119910527.ch3&partnerID=40&md5=54ca6693a112d0c2d274bed31f87e7f2>.

- [48] A. Hanbury, N. Mafirakureva, N. Chicken y L. Bailey, “Evaluating change in a pressured healthcare system: a cross-sectional study of implementation outcomes using routine data indicators and proxies,” *Implementation Science Communications*, vol. 4, n.º 1, 2023, Cited by: 0; All Open Access, Gold Open Access. doi: [10.1186/s43058-023-00471-x](https://doi.org/10.1186/s43058-023-00471-x). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168337946&doi=10.1186%2fs43058-023-00471-x&partnerID=40&md5=eb575908c8abcd7689d4a0648c7a6cf5>.
- [49] P. Gupta, N. Chhikara y A. Panghal, *Harvesting food from weeds*. 2024, 1 – 427, Cited by: 0. doi: [10.1002/9781119793007](https://doi.org/10.1002/9781119793007). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166078921&doi=10.1002%2f9781119793007&partnerID=40&md5=c9daf74a346fb20fd8f838f0c89e108a>.
- [50] U. Farooq, J. Wen, M. I. Tabash y M. Fadoul, “Environmental regulations and capital investment: Does green innovation allow to grow?” *International Review of Economics and Finance*, vol. 89, 878 – 893, 2024, Cited by: 0. doi: [10.1016/j.iref.2023.08.010](https://doi.org/10.1016/j.iref.2023.08.010). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168481672&doi=10.1016%2fj.iref.2023.08.010&partnerID=40&md5=e40a47b0d620677ec8f73fef3d529391>.
- [51] I. Sen, D. Quercia, L. Capra, M. Montecchi y S. Šćepanović, “Insider stories: analyzing internal sustainability efforts of major US companies from online reviews,” *Humanities and Social Sciences Communications*, vol. 10, n.º 1, 2023, Cited by: 0; All Open Access, Gold Open Access, Green Open Access. doi: [10.1057/s41599-023-01672-4](https://doi.org/10.1057/s41599-023-01672-4). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85161422740&doi=10.1057%2fs41599-023-01672-4&partnerID=40&md5=2c82322db611a5b27394285d78daf898>.
- [52] F. Onukwuga et al., “Reflections of maternity service users and midwives’ on the co-creation of interventions to support midwives addressing alcohol during antenatal care,” *Patient Education and Counseling*, vol. 115, 2023, Cited by: 0. doi: [10.1016/j.pec.2023.107896](https://doi.org/10.1016/j.pec.2023.107896). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85166268412&doi=10.1016%2fj.pec.2023.107896&partnerID=40&md5=266b2522ce377c26581901caafa24934>.
- [53] F. Yasmin, M. Khan y Q. Peng, “Optimization of Processing Parameters for 3D Printed Product Using Taguchi Method,” *Computer-Aided Design and Applications*, vol. 21, n.º 2, 281 – 300, 2024, Cited by: 0; All Open Access, Bronze Open Access. doi: [10.14733/cadaps.2024.281-300](https://doi.org/10.14733/cadaps.2024.281-300). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168250976&doi=10.14733%2fcadaps.2024.281-300&partnerID=40&md5=de980461375dbeee73d9386625ef9b18>.
- [54] T. Kripalakshmi y T. Deepa, *Design of DC-DC converters for electric vehicle wireless charging energy storage system*. 2024, 127 – 147, Cited by: 0. doi: [10.1002/9781119792918.ch4](https://doi.org/10.1002/9781119792918.ch4). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168250976&doi=10.14733%2fcadaps.2024.281-300&partnerID=40&md5=de980461375dbeee73d9386625ef9b18>.

- [com/inward/record.uri?eid=2-s2.0-85164855391&doi=10.1002%2f9781119792918.ch4&partnerID=40&md5=0d59703ff2a70f91df528cdc7f6a9168.](https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164855391&doi=10.1002%2f9781119792918.ch4&partnerID=40&md5=0d59703ff2a70f91df528cdc7f6a9168)
- [55] F. Fartout Enayat y M. R. Asgharipour, “Exploring and predicting the biocapacity of various fish farming systems based on modified emergy footprint accounting in the Sistan region of Iran,” *Science of the Total Environment*, vol. 904, 2023, Cited by: 0. doi: [10.1016/j.scitotenv.2023.166195](https://doi.org/10.1016/j.scitotenv.2023.166195). [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168420587&doi=10.1016%2fj.scitotenv.2023.166195&partnerID=40&md5=4fd7ef1da6eae42ff66ff722ac878117>.
- [56] S. U. L. Gabor Schubert, *Scientific Publications on Sustainable Development*, https://www.su.se/polopoly_fs/1.530251.1607009534!/menu/standard/file/sdg-publikationer-2010-2019_gabor_rev3.pdf, 2020.
- [57] Leiden University. “VOSviewer: Visualizing Scientific Landscapes.” (2023), [En línea]. Disponible en: <https://www.vosviewer.com/>.