

Proyecto Espacio de Observación de Inteligencia Artificial en Español

Ámbito 1.2 Portal Estado del Arte

Informe Técnico Año 2

Enrique Amigó, Jorge Carrillo-de-Albornoz, Andrés Fernández, Julio Gonzalo, Guillermo Marco, Roser Morante, Jacobo Pedrosa, Laura Plaza, Eva Sánchez

Natural Language Processing and Information Retrieval Group, UNED

Autor de contacto: Julio Gonzalo - julio@lsi.uned.es

Resumen

En este informe se presenta la Versión 2 del Portal SOTA del proyecto ODESIA, un portal informativo del estado del arte del procesamiento del lenguaje natural en español. El portal se ha desarrollado en la UNED en el marco del proyecto del Espacio de Observación de Inteligencia Artificial en Español, en concreto “Ámbito 1 Estado del arte comparado”, “Actividad 1.2 Portal del estado del arte en español”. Esta versión contiene información sobre 182 tareas y 125 conjuntos de datos procedentes de 95 competiciones de los años 2013 a 2023.

Índice

1. Introducción	2
2. Contenido del portal	3
2.1. Recopilación de datos	3
2.2. Información recopilada	4
2.2.1. Conjunto de datos	6
2.2.2. Tarea de PLN	7
2.2.3. Competición	8
2.2.4. Foro	8
2.2.5. Sistemas de PLN	9
2.3. Perfiles de uso	10
2.3.1. Usuario de consulta	10
2.3.2. Usuario de Administrador	10
2.3.3. Usuario de edición	11
3. Descripción técnica del portal web	11
3.1. Requisitos	11
3.2. Plataformas web disponibles	13
3.2.1. Conclusión	15
3.3. Módulos de Drupal	15
3.4. Lenguajes de programación	16
3.5. Arquitectura	17
3.6. Roles de usuario	18
3.7. Estructura lógica	18
3.7.1. Competición	18
3.7.2. Foro	18
3.7.3. Tema de PLN	18

3.7.4.	Tarea EvALL	19
3.7.5.	Conjunto de datos	19
3.7.6.	Tarea	20
3.7.7.	Resultados	20
3.8.	Visualización del contenido	21
3.8.1.	Estilo	21
3.8.2.	Portada	21
3.8.3.	Tareas	22
3.8.4.	Conjuntos de datos	22
4.	Apéndice: Aspectos técnicos relevantes en el desarrollo de proyectos software	22
4.1.	Mantenibilidad	22
4.2.	ENS: Esquema Nacional de Seguridad	22
4.3.	ENI: Esquema Nacional de Interoperabilidad	23
4.4.	Reglamento General de Protección de Datos	23
4.5.	Informe de técnicas de Search Engine Optimization	23
4.6.	Diseño de la navegabilidad	23
5.	Trabajo Futuro	23
6.	Conclusiones	23
A.	Apéndice: Modelo de base de datos.	25
B.	Apéndice: Páginas web del portal.	26
1.	Introducción	

En este informe se presenta la Versión 2 del portal SOTA del proyecto ODESIA, portal informativo del estado del arte del procesamiento del lenguaje natural (PLN) en español. El portal ha sido desarrollado por el Grupo de Investigación en Procesamiento de Lenguaje Natural y Recuperación de Información de la UNED en el marco del proyecto del Espacio de Observación de Inteligencia Artificial en Español, en concreto Ámbito 1 Estado del arte comparado, Actividad 1.2 Portal del estado del arte en español.

El objetivo del portal es proporcionar información de las tareas de PLN para las cuales se han desarrollado sistemas en español, los resultados obtenidos para estas tareas y los conjuntos de datos existentes para entrenar y evaluar sistemas de PLN. El estudio se centra en el español, puesto que el estudio comparativo español-inglés se realiza en la Actividad 1.1 del Ámbito 1 del proyecto. Esta versión del portal incluye información sobre 182 tareas y 125 conjuntos de datos procedentes de 95 competiciones celebradas en 8 foros científicos diferentes entre los años 2013 y 2023, ambos inclusive.

La información se puede consultar a través del portal web que se entrega juntamente con este informe.¹ Tal y como se describe posteriormente en el documento, el portal proporciona acceso a distintas páginas web y opciones de búsqueda. El portal está diseñado para que diferentes tipos de usuarios, fundamentalmente investigadores, empresas e instituciones, puedan acceder a información sobre el estado del arte y recursos de PLN en español.

Para definir los requisitos de este portal, hemos realizado un análisis previo de las características de recursos existentes como la plataforma Hugging Face² y el sitio web NLP-progress.³ Ambos tienen como objetivo indexar y compartir modelos lingüísticos, conjuntos de datos y herramientas de PLN. NLP-Progress es un repositorio para seguir el progreso del PLN en varios idiomas. Para el español es muy limitado, ya que lista conjuntos de datos para tres tareas, lo que hace que no sea demasiado útil. En

¹<http://portal.odesia.uned.es/>

²<https://huggingface.co/>. Last checked on 20.02.2024.

³<http://nlpprogress.com/>

cambio, Hugging Face es más versátil y diverso. Contiene herramientas de código abierto destinadas a construir, entrenar y desplegar modelos de aprendizaje automático y está dirigido a científicos de datos, investigadores e ingenieros. Abarca una amplia gama de herramientas relacionadas con el PLN, como modelos lingüísticos y conjuntos de datos, así como recursos multimodales. Los usuarios cargan sus propios conjuntos de datos y modelos para su evaluación. Como aspectos positivos, contiene más recursos para el español, 646 conjuntos de datos y 2537 modelos de lenguaje. Sin embargo, como aspectos negativos Hugging Face no permite hacer un seguimiento del SOTA por tarea y no todos los recursos están documentados con el mismo tipo de información, ya que los usuarios determinan cuánta información proporcionan sobre sus productos. La base de datos del Portal que presentamos está diseñada para seguir el estado del arte del español para todas aquellas tareas que se han organizado en competiciones y para almacenar información sistemática sobre los recursos desarrollados para las competiciones.

Varios aspectos hacen que este portal sea novedoso: es el primer portal público que centraliza varios tipos de información relacionada con tareas de PLN en español, incluyendo los resultados del estado del arte - de hecho tampoco conocemos portales similares para otros idiomas; contiene información introducida manualmente; y ofrece múltiples opciones de navegación para facilitar el acceso a la información desde varias perspectivas. La información se ofrece en español e inglés para que los usuarios internacionales puedan beneficiarse de su existencia.

Si bien el presente documento describe el portal tal cual se encuentra en su Versión 2, es importante destacar que, con respecto a la Versión 1, se han realizado las siguientes **mejoras y ampliaciones de funcionalidad**:

- **Multilingüedad:** Se ha implementado la versión en inglés del portal, que en la versión 1 únicamente estaba disponible para el idioma inglés. Esto ha implicado la traducción manual de todos los contenidos.
- Se han implementado nuevas páginas de información para mostrar la información relativa a las competiciones y los foros, respectivamente.
- Se ha implementado un sistema completo de filtros de búsqueda, que permite al usuario refinar sus resultados.
- Se ha actualizado el contenido del portal, incluyendo nuevos foros, competiciones, datasets y tareas celebrados o desarrollados durante el año 2023.
- Se ha llevado a cabo una revisión exhaustiva de los contenidos del portal, con el objetivo de asegurar que no hay información parcial o incompleta.
- Se ha creado un nuevo perfil de usuario, el usuario de edición, con facultad para añadir, eliminar y modificar contenido del portal, pero con menos privilegios que el usuario de administración.

En la Sección 2 del presente documento se describen los datos contenidos en el portal, mientras que en la Sección 3 se describen los aspectos técnicos del desarrollo del mismo. La Sección 5 presenta las líneas futuras de trabajo y mejoras. La Sección 4 recoge aspectos técnicos relevantes en el desarrollo de proyectos software. Finalmente, en la Sección 6 se presentan las conclusiones. Como complemento a este documento, también se entrega un manual de usuario del portal.

2. Contenido del portal

En esta sección se describe la información contenida en la base de datos del portal y el proceso de recopilación de datos.

2.1. Recopilación de datos

Los datos contenidos en el portal se han obtenido mediante revisión manual de las publicaciones de los principales foros o campañas de evaluación ('foros' a partir de ahora por simplicidad de escritura)

en los que se organizan competiciones de PLN. La revisión se ha centrado tanto en foros nacionales como internacionales, partiendo de que las tareas propuestas en estas competiciones constituyen una muestra representativa del estado del arte en PLN. Aunque las competiciones también se pueden organizar al margen de foros, en general, si una competición se propone en un foro, esto garantiza cierto rigor científico, puesto que las tareas pasan por un proceso de selección, la participación en las mismas está abierta a la comunidad internacional de PLN y todos los sistemas participantes se evalúan con los mismos conjuntos de datos y métricas. Además, los conjuntos de datos que se publican en competiciones, suelen seguir usándose posteriormente, de manera que las competiciones definen el panorama del PLN. La información que se ha recopilado se ha extraído de los artículos que describen las competiciones de PLN y/o los conjuntos de datos.

Se ha recopilado información de un período de tiempo que abarca desde 2013 hasta 2023 pues se ha considerado que 2013 era un año representativo para tomarlo como referencia debido a que se publicaron trabajos influyentes que demostraban la eficiencia de los vectores de palabras (word embeddings) en múltiples tareas (Mikolov et al., 2013), cuyo uso causó en su día un cambio en el estado del arte. En general, ir más atrás en el tiempo no añade información relevante, dado que las técnicas de aprendizaje automático utilizadas con anterioridad a la irrupción de los vectores de palabras producen resultados más bajos. No obstante, se ha incluido también alguna competición anterior que se ha considerado relevante por haber publicado conjuntos de datos que han tenido especial repercusión en el avance del estado del arte, como la CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages.

Los foros y competiciones que se han seleccionado porque se han usado datos para el español se enumeran a continuación. Suman un total de 28 ediciones (los años se indican entre paréntesis).

- Cross-Language Evaluation Forum - CLEF (2014, 2017, 2020, 2021, 2022, 2023). <http://www.clef-initiative.eu/>
- Conference on Computational Linguistics - COLING (2022). <https://coling2022.org/coling>
- Computational Natural Language Learning - CoNLL (2009, 2017, 2018). <https://www.signll.org/conll>
- Evaluation of Human Language Technologies for Iberian Languages - IberEVAL (2017, 2018). <https://sites.google.com/view/ibereval-2018>
- Iberian Languages Evaluation Forum - IberLEF (2018, 2019, 2020, 2021, 2022, 2023). <https://sites.google.com/view/iberlef2022/>
- PAN (2018, 2019). <https://pan.webis.de/>
- International Workshop on Semantic Evaluation - Semeval (2014, 2015, 2017, 2018, 2019, 2020, 2022, 2023). <https://semeval.github.io/>
- Second CWI Shared Task (2018). <https://sites.google.com/view/cwisharedtask2018/>

2.2. Información recopilada

El portal incluye información relativa a diversas entidades que tienen una función en el desarrollo del PLN: competiciones, tareas, conjuntos de datos y sistemas. Consultar información sobre estas entidades puede ser de interés para investigadores, empresas, entidades públicas y privadas que financien PLN, y para el público en general, para obtener información relacionada con preguntas como las siguientes: ¿qué competiciones se han organizado para el español?; ¿qué tareas existen para el PLN en español?; ¿qué conjuntos de datos hay disponibles para una determinada tarea, dominio o tipo de texto o variedad lingüística?; ¿qué resultados se han obtenido para una tarea concreta, dominio, tipo de texto o variedad lingüística?.

Respecto a la organización de la información, si bien la mayoría de los repositorios o portales que se presentan como bases de datos del estado del arte o de recursos suelen utilizar como elementos primitivos, en base a los que se organiza la información, los datasets o los sistemas (NLP-progress,⁴ Hugging Face⁵), lo cierto es que, por norma general, un dataset da soporte a varias tareas de PLN. Lo mismo sucede para los sistemas, que son adaptados no a un dataset, sino a una tarea de las varias que pueden contemplarse en un dataset. Por ello, en el portal SOTA se ha considerado que la tarea es la unidad fundamental de información y que un dataset puede estar relacionado con una o varias tareas en función de las anotaciones que tenga. De igual manera, es la tarea la que tiene asociado uno o más sistemas, y no el dataset. Finalmente, la métrica de evaluación depende también del tipo de tarea específica, por lo que las métricas también están asociadas a la tarea. Adicionalmente, para una mejor organización y búsqueda de tareas y datasets, éstos se han asociado a tareas abstractas (clasificación, regresión, etc), que pueden resultar de mayor interés a colectivos empresariales o menos familiarizados con el estado del arte en PLN.

Antes de explicar los datos disponibles para cada entidad es importante definir las entidades y sus relaciones en el contexto de este informe.

Foro: Un foro científico (como CLEF o SemEval) es un evento donde expertos y académicos comparten investigaciones, discuten metodologías y participan en evaluaciones específicas dentro de sus respectivas disciplinas. Estos eventos anuales fomentan la colaboración internacional, la presentación de resultados, y contribuyen significativamente a la innovación y avance tecnológico en sus respectivos campos.

Competición: ('Shared task' en inglés) es un evento que suele tener lugar en el marco de un foro. En una competición se proponen una o más tareas de PLN. Los organizadores de la competición proporcionan los medios científicos necesarios para desarrollar sistemas de PLN para las tareas propuestas, como son la definición de la tarea, los conjuntos de datos textuales con las anotaciones pertinentes y las métricas de evaluación. Los organizadores se encargan de la gestión de la competición, que conlleva hacer publicidad de la misma, recoger y publicar los resultados de los sistemas participantes, y describir todos los aspectos de la competición en un artículo científico. Un ejemplo de competición es "sEXism Identification in Social neTworks" (<http://nlp.uned.es/exist2022/>), organizada dentro del foro de IberLEF 2022.

Tarea de PLN: En el marco de este proyecto, una tarea es una actividad propuesta por los organizadores de una competición con la finalidad de resolver un problema concreto de PLN en el marco de esa competición, que a su vez se organiza en el marco de un foro. Según Schlangen (2021), una tarea establece un mapeo entre un espacio de input y un espacio de output, de los cuales por lo menos uno contiene expresiones en lenguaje natural. El mapeo tiene que adecuarse a la descripción de la tarea. Los organizadores de una competición se encargan de definir las tareas (a veces llamadas subtareas) de la competición y de proporcionar uno o más conjuntos de datos con sus particiones, que los participantes en la tarea de la competición usan para desarrollar sus sistemas. Además, los organizadores proporcionan software de evaluación para evaluar cada tarea de la competición. Los organizadores también tienen que determinar qué métricas se usan para evaluar la tarea, bien sea métricas ya existentes o métricas nuevas que se adecúen mejor a la naturaleza de la tarea. Por tanto, información importante sobre una tarea, además del conjunto de datos y la definición del problema, es la métrica usada para evaluar. En una competición se puede proponer más de una tarea. Por ejemplo, una competición de detección de negación puede dividirse en dos tareas: detección de palabras de negación y detección del alcance de la negación. Cada tarea se puede evaluar por separado y además se puede evaluar la competición de manera global. No obstante, en general, los organizadores de competiciones publican resultados por tarea. Por esta razón, en este portal los resultados de sistemas se asocian a una tarea y no a una competición. Siguiendo con el ejemplo anterior, la

⁴<http://nlpprogress.com/>

⁵<https://huggingface.co/>

competición “sEXism Identification in Social neTworks” proponía dos tareas: identificación de sexismo y categorización de sexismo.

Conjunto de datos: Colección de textos, generalmente con anotaciones. Los organizadores de competiciones de PLN proporcionan conjuntos de datos para cada edición de las competiciones. En general se suele proporcionar un conjunto de datos para todas las tareas de una competición, aunque a veces se proporciona más de uno, dependiendo de si las tareas requieren anotaciones distintas, de la lengua y variedad lingüística que se trate, del tipo de textos, etc. En palabras de [Schlangen \(2021\)](#), mediante el conjunto de datos una tarea se especifica extensionalmente, ya que éste contiene ejemplos del mapeo entre el input y el output que la tarea define. Generalmente los organizadores proporcionan varias particiones del conjunto de datos, una para entrenar, otra para desarrollar y otra para evaluar los sistemas. Por supuesto, existen muchos conjuntos de datos o corpora que no se han publicado para una edición concreta de una tarea. No obstante, como en este proyecto estamos interesados en analizar el estado del arte a partir de los resultados obtenidos en competiciones, en el portal hemos incluido información sobre conjuntos de datos publicados para competiciones. Los datos sobre los conjuntos de datos se han obtenido de los artículos que describen las competiciones.

Sistema de PLN: Un sistema es una implementación concreta de una herramienta o modelo de PLN que resuelve una tarea (o varias). Los participantes en una competición desarrollan un sistema (o varios) para resolver una tarea (o varias). Dado un input, el sistema realiza el mapeo a un output conforme a la descripción de la tarea. Los sistemas son evaluados con el software de evaluación que los organizadores de la competición han proporcionado.

A continuación se describe el tipo de información relativo a cada entidad que se ha incluido en la base de datos del portal.

2.2.1. Conjunto de datos

La información disponible sobre un conjunto de datos determinado es la siguiente:

Nombre del conjunto de datos.

Identificador del conjunto de datos, formado por el nombre y el año de publicación.

Lengua de los textos del conjunto de datos. En las tareas multilingües normalmente se produce un dataset por lengua, de manera que se puedan evaluar los sistemas independientemente en cada una de las lenguas.

Variedad lingüística en la que están escritos los textos. Normalmente los conjuntos de datos contienen sólo una variedad lingüística. Puesto que el español es una lengua hablada en un amplio espacio geográfico, es importante indicar a qué variedad pertenecen los textos del conjunto de datos. En las tareas analizadas hemos encontrado variedades procedentes de los siguientes países: Cuba, Argentina, Chile, Colombia, España, Brazil, Costa Rica, Ecuador, Perú, Uruguay, México. Además hay corpus en Spanglish. En las descripciones de tareas no se suele proporcionar información precisa sobre la variedad lingüística, pues sólo se menciona el país en el que se habla la variedad.

Dominio o área temática de los textos. En función de los datos que hemos encontrado, hemos definido los siguientes dominios: biología, educación, economía, general, salud, legal, noticias, social, turismo, política y cultura.

Tipo de textos. Cada tarea suele centrarse en un tipo de texto concreto. Los tipos pueden referirse a las unidades lingüísticas que se procesan (morfemas, palabras, pares de palabras, frases, textos, etc.), o al formato (tuits, noticias, comentarios de noticias, comentarios de foros, leyes, resúmenes de artículos, etc.).

Tipo de unidades que se usan para cuantificar el tamaño del conjunto de datos, según está especificado en la documentación que lo describe. Pueden ser frases, documentos, tuits, preguntas, lemas, perfiles de autor, etc. A veces el tipo de unidades coincide con el tipo de textos, pero no siempre. Por ejemplo un conjunto de datos pueden contener reseñas de productos, mientras que su tamaño se cuantifica en número de frases.

Número de unidades que conforman el conjunto de datos, correspondiente al tipo de unidades. Esta información procede de la documentación que describe la tarea o el conjunto de datos. En todas las descripciones de conjuntos de datos se menciona algún número de unidades. Idealmente, sería informativo que se proporcionara información sobre el tamaño en función de diferentes unidades, como número de tokens, de frases, de documentos, etc. Algunas descripciones proporcionan esta información, por lo que se han incluido los campos que vienen a continuación:

Número de tokens que conforman el conjunto de datos.

Número de frases que conforman el conjunto de datos.

Número de documentos que conforman el conjunto de datos.

Tamaño digital del conjunto de datos, en términos de MB o GB. Esta información no siempre se proporciona en las descripciones de conjuntos de datos.

Tamaño de las particiones de entrenamiento, desarrollo, evaluación en términos del tipo de unidades que se especifiquen en la descripción del conjunto de datos.

Formato en el que se publica el conjunto de datos: json, texto, columnas separadas por tabuladores, etc.

Anotaciones que se han realizado sobre los datos textuales del conjunto de datos. Las tareas de PLN que se pueden desarrollar con el conjunto de datos dependen del tipo de anotaciones que éste contenga.

Enlace a descripción del conjunto de datos, bien sea una página web o una publicación donde se realiza una descripción detallada del conjuntos de datos.

Tipo de acceso que fundamentalmente puede ser público, con licencia o acceso mediante registro.

Enlace de acceso al conjunto de datos.

Tipo de licencia necesaria para obtener el conjunto de datos.

Publicación científica en la que se describe el dataset.

Enlace a la publicación anterior.

2.2.2. Tarea de PLN

La información recopilada sobre cada tarea de PLN es la siguiente:

Nombre de la competición a la que pertenece la tarea.

Nombre de la tarea.

Descripción de la tarea.

Conjunto de datos que se publica para resolver la tarea.

NLP Topic al que pertenece la tarea.⁶ Una tarea trata un problema concreto, con un conjunto de datos concreto sobre el cual los sistemas participantes obtienen resultados. Las tareas se pueden agrupar en

⁶En PLN se da cierta ambigüedad en el uso del término ‘tarea’, pues éste se usa tanto para referirse a tareas concretas que se proponen con un conjunto de datos concreto, como a tareas en un sentido más abstracto para referirse a grupos de tareas. Para este portal hemos adoptado la definición de tarea de [Schlangen \(2021\)](#), según la cual una tarea establece un mapeo entre un espacio de input y un espacio de output. Llamamos ‘tarea’ a una tarea de PLN concreta para la cual hay sistemas que obtienen resultados, y ‘NLP Topic’ a la etiqueta bajo la cual se puede agrupar ciertas tareas por el tema de PLN que tratan. En los informes técnicos de las competiciones no se reportan resultados por ‘NLP Topic’, sino por tarea. Una alternativa hubiera sido hablar de tareas y subtareas, pero ésta nos parecía más confusa.

áreas o temas (NLP Topic). Los resultados de sistemas no se reportan por área, sino por tarea concreta. Por ejemplo, las tareas de detección de sexismo, clasificación de comentarios ofensivos y detección de agresividad se pueden agrupar dentro del área de detección de odio. Hemos establecido una lista de áreas partiendo de las Áreas de Especialización que la revista *Transaction in Computational Linguistics* proporciona a los revisores para describir sus perfiles. Además, hemos añadido algunas áreas que no constaban. La lista de áreas se proporciona en la Tabla 1.

Idiomas para las que se proporcionan conjuntos de datos en la tarea, codificadas con los códigos ISO 639-1.

Tarea abstracta es el tipo de tarea desde el punto de vista del problema de aprendizaje automático que hay que resolver. En la aplicación web EvaLL⁷ se definen seis tipos de tareas abstractas: classification, sequence labeling, ranking, diversification, clustering, regression, correlation. En este campo se indica a qué tipo pertenece la tarea. Disponer de esta información permite buscar tareas por tipo de problema de aprendizaje automático.

Año de celebración de la tarea.

Enlace a la competición en la que se propone la tarea.

Publicación o referencia bibliográfica de la publicación en la que se describe la tarea.

Enlace a la publicación en la que se describe la tarea.

Enlace al software de evaluación usado para evaluar la tarea.

Métrica oficial usada para establecer el ranking de los sistemas participantes en la tarea.

2.2.3. Competición

La información almacenada sobre cada competición es la siguiente:

Nombre según consta en la publicación que describe la competición.

Descripción detallada de la competición.

Año de edición de la competición.

Foro en el marco del cual se ha organizado la competición.

Enlace a la página web de la competición.

Referencia bibliográfica de la publicación donde se describe la competición.

Enlace a la publicación.

Tareas propuestas en la competición.

2.2.4. Foro

La información almacenada sobre cada foro es la siguiente:

Nombre del foro.

Descripción detallada del foro.

Enlace a la página web del foro.

⁷<http://evall.uned.es/>

(named) entity recognition	paraphrasing
anaphora and co-reference resolution	parsing
argument structure	part-of-speech tagging
automatic speech recognition	pragmatics
bio-medical NLP	processing abbreviations
chatbots	processing events
dialogue systems	processing factuality
discourse processing	processing humor
entity linking	processing negation
fake news detection	question answering
fill mask	relation extraction
hate detection	recommendation systems
image to text	semantic role labeling
information extraction	sentiment analysis
information retrieval	stylistic analysis
language modeling	summarization
lemmatization	text categorization
machine learning for NLP	text classification
machine translation	text generation
morphology	text similarity
multi-word expressions	text simplification
named entity linking	textual entailment
natural language generation	topic modeling
natural language inference	word sense disambiguation
normalization	

Tabla 1: Áreas de PLN.

2.2.5. Sistemas de PLN

La recogida de información relativa a los sistemas que han participado en las competiciones está enfocada fundamentalmente a recopilar los resultados que se han obtenido para informar sobre el estado del arte. La información se ha obtenido de los artículos donde los organizadores describen las competiciones, las tareas y los resultados obtenidos por los sistemas participantes. Para cada tarea, hemos recopilado los resultados de, como mínimo, los cinco mejores sistemas, considerando que estos son suficientemente representativos del estado del arte en el momento que tiene lugar la competición.

La información almacenada sobre cada sistema es la siguiente:

Nombre del sistema.

Competición en la que participa.

Tarea en la que participa.

Track en el que participa, si hubiera tracks. Un track es una versión de la tarea en la que se varían algunos aspectos, como puede ser el conjunto de datos o los procesos de procesamiento permitidos. Cada track se evalúa independientemente. La mayoría de tareas no tienen tracks.

Año de edición de la tarea.

Métrica Oficial utilizada para confeccionar el ranking de la competición.

Resultados obtenidos por el sistema para las métricas utilizada para realizar el ranking en la competición.

Publicación. Referencia bibliográfica de la publicación en la que se describe la tarea, y de la cual se han obtenido los resultados que se introducen en el portal.

Enlace a la publicación.

2.3. Perfiles de uso

La información descrita en las secciones anteriores puede ser usada por varios tipos de usuarios. Distinguiendo entre el usuario administrador, de consulta y el usuario de edición.

2.3.1. Usuario de consulta

El usuario de consulta en el portal del estado del arte generalmente busca acceder a información actualizada y relevante sobre los avances y desarrollos más recientes en el campo del procesamiento del lenguaje. Dentro del usuario de consulta, a su vez, nos encontramos con diferentes perfiles:

- Investigadores que buscan información sobre:
 - Las tareas de PLN que existen en español.
 - Los resultados que se obtienen para una tarea.
 - Conjuntos de datos existentes por tarea, dominio, variedad lingüística de los textos, tipo de textos, anotaciones.
 - Tareas existentes por dominio, variedad lingüística, tipo de textos, anotaciones.
 - Cómo acceder a conjuntos de datos.
 - Medidas de evaluación utilizadas en general y por tareas.
- Empresas que quieren acceso rápido a información sobre:
 - Variedad de tareas de PLN que existen.
 - Los resultados que se obtienen para una tarea.
 - Conjuntos de datos existentes por tarea y anotaciones que contienen.
 - Cómo acceder a los conjuntos de datos.
- Entidades que financian actividades relacionadas con el PLN y necesitan:
 - Detectar vacíos en áreas de PLN en español.
 - Evaluar la originalidad de propuestas de proyectos en relación a las tareas existentes en español.
 - Contextualizar los resultados de proyectos en el estado del arte del PLN en español.
 - Analizar cambios en el estado del PLN en español a lo largo del tiempo.
- Ciudadanos interesados en información sobre el PLN en español.

2.3.2. Usuario de Administrador

El usuario administrador de este proyecto, es el responsable de mantener y gestionar eficazmente el sitio web, asegurando su correcto funcionamiento y adaptándolo a las necesidades específicas del proyecto. Este usuario tiene acceso completo y control total sobre la administración del sitio web. Al crear un nuevo sitio en Drupal, se establece un usuario administrador durante el proceso de instalación.

El usuario administrador tiene la capacidad de:

- Gestionar y configurar todos los aspectos del sitio web, incluyendo la estructura del contenido, la apariencia, las configuraciones del sitio y los permisos de usuario.
- Crear, editar y eliminar contenido de cualquier tipo, como artículos, páginas, comentarios, etc.
- Instalar, activar, desactivar y configurar módulos y temas para extender las funcionalidades y personalizar el diseño del sitio.

- Gestionar usuarios y asignar roles y permisos a otros usuarios del sitio.
- Realizar tareas de mantenimiento, como realizar copias de seguridad, actualizar el núcleo de Drupal y los módulos, y solucionar problemas de rendimiento o seguridad.

2.3.3. Usuario de edición

Debido al amplio alcance y poder del usuario administrador y de los riesgos que esto conlleva para la seguridad del sitio, se ha decidido restringir el acceso de este usuario a sólo una cuenta para evitar posibles riesgos de seguridad y crear un rol específico para la gestión de contenido. Este rol se define como usuario de edición.

El usuario de edición no solo consume la información, sino que también contribuye a la actualización y mejora del contenido. Puede ser un experto, investigador o académico con conocimientos especializados en el área que desea añadir o corregir información, o puede ser un usuario con perfil administrativo que canalice las peticiones de la comunidad investigadora. De esta manera, se consigue el mantenimiento de un recurso dinámico, relevante y en constante evolución.

Puesto que los usuarios potenciales quizás no estén familiarizados con las opciones de acceso a información que el portal ofrece, como parte del proyecto se ha elaborado un manual de usuario detallado en el que se muestra cómo buscar información desde diversas perspectivas, así como editar (dar de alta, baja y modificar) la información contenida en el portal. El manual se entrega en un documento aparte.

3. Descripción técnica del portal web

En la sección anterior hemos descrito el tipo de información que contiene el portal. En esta sección nos centramos en describir los aspectos técnicos del mismo partiendo de los requisitos y necesidades básicos que deberá satisfacer y valorando diferentes tecnologías con las que desarrollar el proyecto. Finalmente, se presentan las herramientas, frameworks, lenguajes y librerías que se han utilizado para la implementación del proyecto, así como su estructura web, la organización interna de contenidos y de la base de datos.

3.1. Requisitos

Para definir los requisitos de este portal se han analizado previamente las características de otros portales existentes como son Hugging Face (<https://huggingface.co/>) y NLP-progress (nlpprogress.com). Ambos son plataformas que pretenden indexar y compartir modelos de lenguaje, conjuntos de datos, herramientas e información del estado del arte en PLN, siendo Hugging Face más transversal y NLP-Progress más centrado en el ámbito del PLN.

A partir del análisis de estos portales de referencia se han definido unos requisitos básicos:

- El portal debe ser administrable. Pese a ser un portal público de muestra de contenidos que no requiere suscripción, es necesario que disponga de un panel que permita administrar fácilmente tanto los contenidos del mismo, como la visibilidad de estos. Además, para que este proyecto perdure en el tiempo es requisito que sea fácilmente administrable y ampliable sin depender exclusivamente de conocimientos en programación.
- En el desarrollo de un proyecto, se pueden encontrar principalmente 3 tipos de estructura:
 - Sistema integrado en el que todo el código se desarrolla sin separación, lo cual produce como principal desventaja que un error o cambio en un aspecto independiente del proyecto puede afectar a otras partes del mismo.
 - Sistema con extensiones en el que el código suele estar formado por un núcleo o *core* y a éste se le pueden añadir extensiones que añaden o modifican funcionalidades de este núcleo a partir de una API. La ventaja de este aspecto, es que diferentes desarrolladores pueden modificar el comportamiento del proyecto mediante estas extensiones sin afectar el desarrollo de otros desarrolladores. Como principal desventaja, las extensiones no suelen tener comunicación entre ellas, por lo que para poder implementar alguna funcionalidad conjunta entre diferentes extensiones se tiene que recurrir a una tercera implementación personalizada.

- Sistema modular que es un formato similar al anterior, con la diferencia de que en lugar de extensiones, se trabaja con módulos que pueden estar interconectados entre ellos, con la clara ventaja de que se pueden unir desarrollos independientes para una funcionalidad concreta.

Para este proyecto se ha optado por construir el portal en una estructura modular, de modo que se puedan ampliar fácilmente las funcionalidades actuales si el proyecto así lo requiere. Además, la modularidad posibilita que el mantenimiento sea más fácil y transferible a otras personas en el futuro.

- **Multilingüedad (EN/ES).** Dado que el proyecto en global está orientado a comparar la brecha del desarrollo del PLN entre Inglés y Español, es necesario que los usuarios puedan acceder a la información en las dos lenguas. Por esta razón se requiere que el portal cumpla con los estándares de internacionalización y que permita, no solo gestionar estos dos idiomas, sino que se pueda gestionar un tercero o más, si en un futuro se considera necesario, por ejemplo ampliándolo a otras lenguas del estado. Además es importante que el portal tenga proyección internacional, razón por la cual debe ofrecerse en inglés.
- **Base de datos centralizada.** Puesto que dentro del proyecto del Observatorio, además de este portal se desarrollará un leaderboard y una plataforma de evaluación (EvALL), se plantea como requisito básico establecer una base de datos común para los 3 subproyectos, de modo que se puedan retroalimentar entre ellos. Una base de datos centralizada permite a los 3 proyectos interactuar entre ellos y que la información aparezca siempre actualizada en los tres. Además, de este modo se reducen los errores de concurrencia que puedan darse entre los proyectos.
- **Virtualización.** Para una gestión fácil del proyecto, se requiere que este esté creado sobre una estructura de contenedores de software de virtualización como Docker. Docker es una plataforma de software que genera independencia entre el entorno en que se ejecuta el software y el entorno físico reduciendo de este modo los problemas de compatibilidad de software. Además ofrece ventajas como portabilidad, velocidad de implantación y aislamiento de otros proyectos dentro del mismo servidor optimizando al máximo el uso de los recursos disponibles.
- **Diseño multiplataforma.** Para maximizar la accesibilidad de la plataforma, se pretende que ésta esté basada en un diseño responsivo, es decir que automáticamente adapte la interficie de usuario al dispositivo con el que se está visualizando no solo en termino de escalado sino incluso en el propio diseño o en el comportamiento de sus componentes, por ejemplo mostrando un menu desplegable en pantallas pequeñas o que el numero de componentes en una lista sea flexible en función del tamaño de la pantalla.
- **Buscador.** Con el fin de facilitar las búsquedas de contenidos a los usuarios, se define como requisito imprescindible que el portal tenga un buscador, mediante el cual el usuario pueda encontrar contenidos de diferentes tipos con una única búsqueda.
- **Basado en Software Libre.** Los proyectos basados en este tipo de licencias están conformados principalmente por 4 libertades que deben cumplir:
 - La libertad de ejecutar el programa como se desee, con cualquier propósito (Libertad 0).
 - La libertad de estudiar cómo funciona el programa, y cambiarlo para que haga lo que se desee (Libertad 1). El acceso al código fuente es una condición necesaria para ello.
 - La libertad de redistribuir copias para ayudar a otros (Libertad 2).
 - La libertad de distribuir copias de sus versiones modificadas a terceros (Libertad 3). Esto le permite ofrecer a toda la comunidad la oportunidad de beneficiarse de las modificaciones. El acceso al código fuente es una condición necesaria para ello.

Estas características o libertades causan que los proyectos de software de mayor calidad se vean apoyados por comunidades de desarrolladores que producen finalmente software de gran calidad

muy revisado y con actualizaciones continuas con las que se mejoran aspectos como la seguridad o la velocidad. Además facilitan que el código desarrollado se libere como código abierto.

3.2. Plataformas web disponibles

Para este proyecto se valoraron inicialmente 3 plataformas de desarrollo de webs, todas ellas basadas en formatos de código abierto y que cumplen en mayor o menor medida las especificaciones básicas especificadas en la Sección 3.1. Describimos a continuación las tres opciones con sus ventajas e inconvenientes.

Wordpress (<https://wordpress.com>) es el gestor de contenidos más popular del mundo actualmente. Según las cifras publicadas en su web, el 40 % de las páginas web publicadas en Internet están hechas con Wordpress.

Sus principales **ventajas** son:

- Sencillez de uso. Es un gestor simple de gestionar una vez se conocen las bases de creación y gestión de contenido, tanto en contenido como en administración.
- Actualización constante. Las actualizaciones son prácticamente diarias, ya sea en su motor principal o de sus múltiples plugins.
- Gran cantidad de plugins y temas. Wordpress cuenta con (quizás) la mayor comunidad de desarrolladores en un proyecto de sus características. Mientras que en otros proyectos similares se puede encontrar una solución genérica para un problema determinado, en Wordpress podemos encontrar normalmente al menos 3 soluciones más concretas para el mismo problema.
- Orientación a SEO. Quizás una de las máximas de este gestor de contenidos es su capacidad de orientación a SEO junto con los plugins freemium Yoast o AllInOneSEO, característica que ha ayudado enormemente a impulsar este gestor de contenidos.

Principales **inconvenientes**:

- Temas y plugins de pago. Pese a contar con una gran librería de temas y plugins, la mayor parte de éstos son en formato freemium/suscripción o licencia, por lo que en varios casos su funcionalidad principal para este proyecto requerirá de la contratación de un servicio de pago periódico que imposibilita el uso de estas herramientas para este proyecto.
- Orientación a blog. Pese a que Wordpress en su base es un gestor de contenidos, está orientado principalmente a la creación de blogs y/o páginas web de contenidos. Por lo que cualquier funcionalidad que se se salga de este aspecto, requiere la instalación de extensiones concretas que añadan dicha funcionalidad.
- Multiidioma. Aunque de manera nativa se puede instalar Wordpress en cualquier idioma, está orientado a trabajar con solo uno. Este aspecto difiere de lo que se refiere a la gestión de contenidos en diferentes idiomas. En este punto el plugin más popular y que mejor trabaja este aspecto, es, sin duda, WPML (<https://wpml.org/es/>). La desventaja de este plugin es que es de pago desde el inicio, aunque con licencia de por vida.
- Seguridad. Si bien es cierto que este gestor tiene fama de inseguro, esta fama le viene dada principalmente por dos causas. La primera porque entre los millones de páginas web desarrolladas con él, son muchas las que no están mantenidas y actualizadas convenientemente. A pesar de que su proceso de actualización es sencillo, e incluso automático en el core de las versiones 4+, es imprescindible tener actualizados tanto temas como plugins e incluso aquellos que se tienen desactivados. La segunda causa es que dada la gran cantidad de páginas realizadas en este gestor, son también muchos los ojos puestos en sus vulnerabilidades, y existen múltiples bots encargados de rastrear y localizar versiones de Wordpress desactualizadas con según qué configuraciones vulnerables.
- Estructura basada en plugins independientes. Pese a ser una de las opciones más flexibles que hay, una de las características de Wordpress para la implementación de este proyecto es que

gran parte del desarrollo de la estructura básica se debería hacer por código en php, lo cual provocará que para realizar futuras mejoras o modificaciones (sencillas) sea necesario disponer de conocimientos sobre este entorno.

Gestor de contenidos Drupal (<https://www.drupal.org/>) es otro de los gestores de contenidos más populares hoy en día. Pese a que su cuota es muy inferior a la de Wordpress, alrededor de un 2-3 %, en general, se pueden encontrar proyectos de alta calidad desarrollados en este gestor de contenidos. En su portal, se pueden encontrar gran cantidad de ejemplos de casos de éxito (<https://www.drupal.org/case-studies>) relevantes entre los que destacan:

- <https://www.nasa.gov/>
- <https://www.reaganlibrary.gov/>
- <https://www.gsb.stanford.edu/>
- <https://spia.princeton.edu/>

Sus principales **ventajas** son:

- **Framework.** En sus últimas versiones se considera que Drupal está más cerca de un Framework (basado en <https://symfony.es/>) de desarrollo que de un gestor de contenidos. Es por ello que se recomienda para proyectos de gran alcance y que requieran una personalización determinada en cuanto a funcionamiento.
- **Tipos de contenido y vistas.** Como principal ventaja para este proyecto, destacan dos módulos inherentes a este sistema, el de gestión de contenidos y el de vistas, con los que por un lado se puede definir casi cualquier tipo de contenido con sus campos y por otro crear diferentes tipos de listados de contenido adaptado a las necesidades del proyecto. Estos módulos ahorran escribir código, a la vez que en cualquier momento se puede añadir código para personalizar al máximo la estructura generada. Pese a que en las primeras versiones de Drupal esta funcionalidad estaba separada en dos módulos independientes, son quizás los dos módulos de mayor recorrido de la comunidad, con gran cantidad de extensiones como la posibilidad de incluir formularios-embedidos, lo que permite la creación de entidades dentro de entidades relacionadas.
- **Estructura modular.** A diferencia de Wordpress, Drupal trabaja en una estructura modular donde cada módulo puede depender o ser requisito para otros, por lo que se facilita la escalabilidad y el desarrollo de nuevas funcionalidades.
- **Gestión multidioma integrada.** Una de las grandes ventajas de Drupal para este proyecto es su gestión de contenidos multidioma integrados y aplicables a todos los tipos de contenido que se desarrollen.
- **Desarrollo en línea.** Otra de las características más destacables de Drupal es que está concebido para que gran parte del desarrollo e implementación de soluciones se pueda hacer sin tener que desarrollar código directamente, sino que se pueden hacer pequeñas variaciones y ampliaciones directamente desde su panel de control.
- **Gran capacidad de personalización.** Ya sea con los módulos inherentes como con módulos de desarrollo en bloques, Drupal permite configurar hasta el más mínimo detalle tanto su área pública como en la privada e incluso diferentes configuraciones en función del rol o el usuario concreto.

Las posibles desventajas son:

- Una de las principales desventajas que se conoce de este CMS es su poca compatibilidad de módulos con versiones anteriores. Es un problema que destaca sobretodo en versiones anteriores a la 7. Pese a ello, en un estudio previo del proyecto no se ha encontrado ningún módulo de estas versiones que sea requerido, y en todo caso se puede plantear la posibilidad de un desarrollo ad-hoc que pueda suplir una necesidad concreta que no se encuentre entre el listado de módulos compatibles.

- Está basado en un lenguaje de programación precompilado, lo cual puede afectar en situaciones de gran carga de visitas. Esto se soluciona con sistemas de *cache* internos y externos.
- Al estar programado en sistema modulares, el desarrollo de tests unitarios tiene que ser realizado para cada módulo de manera independiente y no como un sistema integrado, lo cual puede dificultar el desarrollo continuo.
- Uno de los hitos destacables en Drupal, respecto a Wordpress como Gestor de Contenidos, es su pronunciada curva de aprendizaje inicial, que requiere un tiempo para acostumbrarse tanto a la administración global y por módulos, como también a la abstracción de base de datos que se realiza por nodos y campos. Esto además complica el trabajo directo desde base de datos.

Django Framework/CMS o Flask La primera intención con Django fue la de crear un gestor de contenidos modular desarrollado en Python. Como gestor de contenidos, tiene una estructura básica bastante fiable, que puede cumplir con los requisitos básicos del proyecto.

Sus principales **ventajas** son:

- Está basado en Django y Python. Al ser un CMS basado en Django, importa todas las ventajas de este framework.
- Multiidioma. Igual que en los casos anteriores, tiene implementada una solución nativa para la gestión de lenguaje a nivel de textos básicos, pero requiere de un desarrollo concreto para esta gestión básica.
- Trabajar directamente sobre un framework permite alcanzar más en detalle el desarrollo a medida en funcionalidades muy concretas e incluso poder definir estructura cliente-servidor que entornos de Gestor de contenidos viene preestablecida.
- El lenguaje de programación Python es uno de los más populares actualmente y es el mismo que en el que se va a desarrollar la librería de evaluación EvALL por lo que trabajar en un mismo lenguaje puede ser una ventaja.

Las posibles desventajas son:

- El hecho de trabajar con un framework obliga a desarrollar muchas funcionalidades transversales que en un gestor de contenidos ya vienen implementadas por defecto, como por ejemplo una administración o sistemas de autenticación.
- Las actualizaciones en este formato vienen dadas del framework en si y de librerías añadidas, pero no por funcionalidad, por lo que en un entorno de código abierto sin actualizaciones recurrentes se pueden generar graves vulnerabilidades de seguridad.
- Pese a ser multiidioma, es necesario implementar colecciones de traducciones que en muchos casos ya se encuentran implementadas por la comunidad en los casos de Wordpress y Drupal.

3.2.1. Conclusión

Tras estas valoraciones, para este proyecto se opta por una estructura híbrida entre el gestor de contenidos Drupal que abarca la mayor parte de las necesidades del proyecto en general y la implementación de una API Restindependiente con Flask/Python que permita suplir las necesidades que el gestor no alcance.

3.3. Módulos de Drupal

Además de lo descrito anteriormente, como principales beneficios del CMS Drupal observamos que se incluyen diferentes **módulos** útiles para el desarrollo de este proyecto, tanto producidos por la comunidad, como incluidos en el propio gestor, Algunos ejemplos son los siguientes:

- CCK (Content Creation Kit). Desde los inicios de Drupal, este ha sido uno de los módulos más populares desarrollados por la comunidad, dado que permite gestionar los tipos de entidades con las que se va a trabajar sin necesidad de desarrollo directo. Actualmente se encuentra incluido de manera nativa en Drupal.

- **Views.** Este módulo al igual que el anterior, es muy popular y se incluye de manera nativa en Drupal. Gracias a él se pueden crear páginas, bloques y secciones con listados de contenidos sin necesidad de acceder directamente a la base de datos. Cabe destacar que estos dos módulos producen una abstracción de la base de datos de tal manera que se evitan errores y facilitan el trabajo en sistemas basados en MVC.
- **Taxonomy manager.** Las taxonomías o vocabularios son un tipo de contenido específico dentro de Drupal que permiten crear listados de catalogación que pueden ser usados dentro del resto de contenidos. De este modo, se pueden crear etiquetas, menús, secciones, etc., con las que facilitar la navegación y la búsqueda de información al usuario. Este módulo proporciona una interfaz para administrar taxonomías dentro de Drupal, proporcionando al administrador las siguientes operaciones y características clave:
 - Gestión en vista de árbol dinámica.
 - Borrado masivo.
 - Adición masiva de nuevos términos.
 - Gestión de términos en jerarquías.
 - Fusión de términos.
 - Ordenación de las listas de taxonomía.
 - Formulario de edición de términos con tecnología AJAX.
 - Interfaz de búsqueda sencilla.
 - CSV Exportación de términos.
 - Compatibilidad con i18n para vocabularios multilingües (términos por idioma).
 - Interfaz de doble árbol para mover términos en jerarquías, agregar nuevas traducciones y cambiar términos entre diferentes vocabularios.
- **Backward Compatibility.** Este módulo permite a Drupal trabajar con módulos de versiones anteriores del Gestor de contenidos con las versiones más recientes (9+, 10+).
- **CSV Importer.** Módulo que facilita la gestión de contenido en Drupal permitiendo importarlo y exportarlo desde y a formato csv.
- **Better Exposed Filters.** Reemplaza los formularios de filtros predeterminados de las vistas de Drupal con campos más específicos para mejorar la experiencia del usuario. Cuando expone un filtro, permite que el usuario interactúe con la vista, lo que facilita la creación de una búsqueda avanzada personalizada.
- **Views Bootstrap.** Este módulo integra los formatos de listado existentes en el framework Bootstrap en el módulos de gestión de vistas Drupal.

Una de las principales características de Drupal, es su capacidad de trabajar en **sistema de capas separadas (MVC)**, donde todo el entorno visual del proyecto se implementa mediante templates (plantillas) hereditarias entre si. Es decir, que una plantilla puede heredar funcionalidades de otra.

De este modo, para el desarrollo de este proyecto se ha utilizado la plantilla oficial de Bootstrap a partir de la cual se ha implementado una plantilla que hereda sus principales componentes.

3.4. Lenguajes de programación

Tal como se ha comentado en puntos anteriores, se ha optado por utilizar Drupal como base para la implementación de este proyecto. Este, en sus versiones más recientes (8,9 y 10) está basado a su vez en el Framework Symphony (<https://symfony.es/>), el cual está desarrollado en PHP, lenguaje que actualmente abarca cerca del 85 % de los sitios web en Internet.

Como base de datos, se ha optado por MariaDB, un Sistema Gestor de Base de datos heredero de Open Source de MySQL, el cual en sus versiones más recientes se ha propietarioizado.

Por esta razón, la implementación y adaptación de Drupal al proyecto para añadir o modificar funcionalidades se debe hacer principalmente con estos dos lenguajes. Además, se han utilizado otras tecnologías y/o lenguajes para el alcance total de este proyecto.

Motor de plantillas Twig Twig es un motor de plantillas desarrollado para el lenguaje de programación PHP y que nace con el objetivo de facilitar el trabajo relativo a las vistas a los desarrolladores de aplicaciones web que utilizan la arquitectura MVC, gracias a que se trata de un sistema que resulta muy sencillo de aprender y que es capaz de generar plantillas con un código preciso y fácil de leer.

Una plantilla es un archivo de texto que puede arrojar resultados en formatos como HTML, XML, CSV, etc., y que está formado por expresiones de control y variables, las cuales serán reemplazadas por valores una vez que la plantilla sea evaluada. Cuando nos referimos al lenguaje PHP, una de las plantillas más utilizadas es una plantilla PHP, en la que se mezcla texto interpretado por PHP y en el que se mezclan etiquetas HTML y código PHP para formar la vista que verá el usuario.

CSS Para la implementación del diseño en ámbito web, se trabaja con CSS (Cascade Style Sheet), que es un metalenguaje que permite definir aspectos de diseño al HTML mediante reglas de diseño como color, tamaño, etc.

Para este proyecto, se ha optado por una ampliación de CSS, el framework SASS, que es un preprocesador CSS, una herramienta que nos permite generar hojas de estilo de manera automática, añadiéndoles características que no tiene CSS y que son propias de los lenguajes de programación, como pueden ser variables, funciones, selectores anidados, herencia, etc.

Estas características de los procesadores nos permiten, además, que el CSS que se genera sea más fácil de mantener y más reutilizable.

JavaScript JavaScript es un lenguaje de programación que a diferencia de los demás, se puede ejecutar directamente en el navegador, otorgando herramientas para crear animaciones, ejecuciones en segundo plano, trabajar con cookies y/o sesiones, etc. Además de las librerías y funcionalidades incluidas, se utiliza este lenguaje para mejorar la experiencia de usuario. También se han incluido diferentes librerías populares en JavaScript, como es jQuery.

Bootstrap es uno de los framework de frontend más populares, desarrollado por Twitter y publicado con licencia Open Source concretamente Licencia MIT (<https://getbootstrap.esdocs.com/docs/5.1/about/license/>). Este framework además de incluir muchas herramientas para maquetar el diseño de una web, está orientado a desarrollo responsivo, es decir, que el diseño se adapta automáticamente a la resolución del dispositivo con el que se está navegando en la página, incluso en pantallas de menor tamaño, ya sea móvil o tablet.

3.5. Arquitectura

Tal como se ha comentado anteriormente, para la arquitectura de este proyecto se parte de la base de trabajar en contenedores aislados implementados en Docker, de tal manera que cada parte del proyecto pueda trabajar de manera independientemente para evitar problemas de concurrencia de software, aunque al mismo tiempo estos contenedores trabajan de forma unificada en un entorno de red virtualizada por el propio sistema.

Dentro de este sistema podemos encontrar diferentes redes virtualizadas de las que destacan principalmente una privada, a la que solo tienen acceso los contenedores que así se hayan configurado, y una red pública conectada con el entorno físico donde se ejecuta el sistema. Si a este se le da salida, se puede acceder desde Internet. De este modo, se garantiza una comunicación eficiente entre los diferentes contenedores y se maximiza la seguridad en aspectos como la base de datos y/o la API.

Los contenedores se describen a continuación:

- Contenedor Apache/PHP/Drupal que contiene el sistema de archivos con la lógica de la programación del sistema. Es la parte central que gestiona los otros nodos del sistema. Este contenedor tiene doble

conexión, una dentro del propio entorno virtualizado (privada) y otra a la red pública, dado que por un lado tiene que trabajar con los diferentes contenedores y por otro tiene que ser accesible a los usuarios.

- Contenedor MariaDB, que contiene el SGBD MariaDB y la base de datos del proyecto. Este contenedor únicamente es accesible dentro de la red privada.

3.6. Roles de usuario

Los usuarios de este portal desde el punto de vista de los casos de uso se han descrito en la Sección 2.3. Fundamentalmente se trata de investigadores, empresas y entidades de financiación. No obstante, desde el punto de vista técnico se definen dos tipos de usuario:

1. Anónimo: usuario anónimo dentro del portal que inicialmente no precisa de registro y/o inicio de sesión para satisfacer sus necesidades de información.
2. Editor: Usuario registrado con permisos para añadir, borrar y modificar contenidos del portal.
3. Administrador: el portal debe contar con al menos un perfil de administrador que permita gestionar los contenidos de manera eficiente.

3.7. Estructura lógica

A partir de la especificación hecha en la Sección 2.2, se ha creado una estructura lógica de los contenidos del portal dentro de la abstracción de la base de datos que genera Drupal con su módulo CCK de definición de tipos de contenido. De este modo, se han generado los tipos de contenido que se detallan a continuación, con sus atributos y especificación de cada campo, indicando cómo se almacena dentro de la abstracción que hace Drupal y la representación del mismo en los formularios de creación de contenido. En la Figura 2 se muestra un esquema de la abstracción de la base de datos generada dentro de Drupal mediante los tipos de contenido, que describimos a continuación.

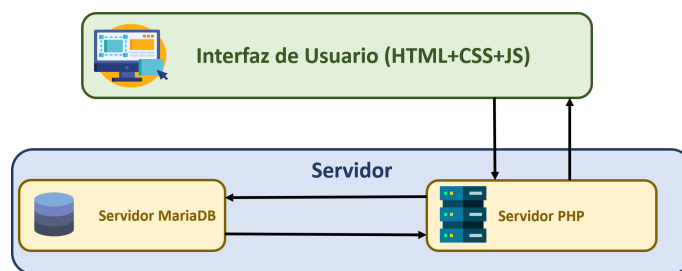


Figura 1: Arquitectura funcional del Proyecto.

3.7.1. Competición

Este tipo de contenido está creado como una taxonomía (descritas en el apartado 3.3). Inicialmente no es necesario añadir más atributos que el nombre, el foro y la descripción de competición, ya que el resto de información se extrae de los contenidos relacionados con ésta en la base de datos, es decir, Foro y Tareas.

3.7.2. Foro

Este tipo de contenido está creado también como una taxonomía. Inicialmente no es necesario añadir más atributos que el nombre, ya que el resto de información se extrae de los contenidos relacionados con Foro en la base de datos, es decir, Competición.

3.7.3. Tema de PLN

Este tipo de contenido está creado también como una taxonomía. Inicialmente no es necesario añadir más atributos que el nombre, ya que el resto de información se extrae de los contenidos relacionados.

3.7.4. Tarea EvALL

Este tipo de contenido está creado también como una taxonomía. Inicialmente no es necesario añadir más atributos que el nombre, ya que el resto de información se extrae de los contenidos relacionados.

3.7.5. Conjunto de datos

Este tipo de contenido permite almacenar entidades de tipo conjunto de datos con sus atributos principales:

- ID. Identificador del conjunto de datos. Tipo: Texto sin formato.
- Nombre público del conjunto de datos. Tipo: Texto sin formato.
- Descripción. Descripción del dataset. Tipo: Texto sin formato.
- Enlace descripción Dataset. Enlace externo donde se puede encontrar la descripción del conjunto de datos. Tipo: URL.
- Idioma. Idioma o idiomas, en caso de ser un conjunto de datos en varios idiomas, codificados siguiendo el estándar ISO-639-1. Tipo: Lista de texto.
- Año de creación del conjunto de datos. Tipo: Número entero.
- Dominio de los datos del conjunto de datos. Tipo: Lista de texto.
- Formato de los datos (csv, pdf, etc). Tipo: Lista de texto.
- Tipo de textos que conforman el conjunto de datos, legal, médico, etc. Tipo: Lista de texto.
- Instancias. Número de instancias totales que se pueden encontrar dentro del conjunto de datos. Tipo: Número entero.
- Unidades. Tipo de elementos de los que está formado el conjunto de datos: documentos, tuits, palabras, etc. Tipo: Lista de texto.
- Tokens. Cantidad de tokens contenidos en el conjunto de datos. Tipo: Número entero.
- Frases. Cantidad de frases contenidas en el conjunto de datos. Tipo: Número entero.
- Docs. Cantidad de documentos que conforman el conjunto de datos. Tipo: Número entero.
- Tamaño en MB del conjunto de datos. Tipo: Número entero.
- Set splits. Tamaño de las particiones del conjunto de datos.
 - Tamaño TrainSet. Tipo: Número entero.
 - Tamaño DevSet. Tipo: Número entero.
 - Tamaño TestSet. Tipo: Número entero.
- Información adicional tamaño. Espacio en el que se puede añadir una descripción adicional sobre el tamaño del conjunto de datos o sobre las particiones. Tipo: Espacio de texto.
- Anotación etiquetado. Espacio de texto simple en el que se puede especificar el tipo de anotaciones contenidas en el dataset. Tipo: Texto sin formato.
- Enlace guía anotación. Enlace al documento en el que se describen las anotaciones. Tipo: URL.
- Acceso. Lista desplegable para indicar como es el acceso al conjunto de datos (público, privado, con registro, etc.). Tipo: Lista de texto.

- Enlace acceso a datos. Enlace a la página donde se puede obtener el conjunto de datos. Tipo: URL.
- Publicación. Espacio en el que se incluye la referencia bibliográfica de la publicación donde se describe el conjunto de datos. Tipo: Texto largo sin formato.
- Enlace publicación. Campo con el enlace a la publicación relacionada al conjunto de datos. Tipo: URL.
- Licencia. Licencia con la que se publica el conjunto de datos. Tipo: Lista de texto.
- Notas. Espacio en el que se pueden añadir notas relacionadas al conjunto de datos. Tipo: texto sin formato; múltiple.

3.7.6. Tarea

- Título. Campo que incluye el título de la tarea. Tipo: Texto simple.
- Competición. Nombre de la competición. Tipo: Término de taxonomía.
- NLP Topic. Área de PLN a la que pertenece la tarea. Tipo: Término de taxonomía.
- Tarea EvALL. Tipo de tarea desde el punto de vista de la evaluación. Tipo: Término de taxonomía.
- Idiomas Tarea. Idioma o idiomas, en caso de ser un conjunto de datos en varios idiomas, codificados siguiendo el estándar ISO-639-1. Tipo: Lista de texto.
- Descripción de la tarea. Campo en el que se puede añadir una descripción de la tarea. Tipo: Texto con formato.
- Año. Año de edición de la tarea. Tipo: Número entero.
- Publicación. Referencia bibliográfica de la publicación donde se describe la tarea. Tipo: Texto sin formato.
- Enlace publicación. Campo en el que se puede añadir un enlace para encontrar la publicación asociada a la tarea. Tipo: URL.

3.7.7. Resultados

Este tipo de contenido hace referencia a los resultados obtenidos por los sistemas al resolver una tarea. Los resultados están vinculados a una tarea. Una tarea puede estar evaluada con más de una métrica y tener múltiples resultados asociados por métrica.

- Identificador. Identificador único del resultado que permite indexar éste en base de datos y relacionarlo con los otros tipos de contenido. Tipo: Texto simple.
- Tarea. Tarea en la que se han obtenido estos resultados. Tipo: Relación a contenido de tipo tarea.
- Año. Año en que se han obtenido los resultados obtenidos. Tipo: Número.
- Track. Track de la tarea en la que se han obtenido los resultados. Tipo: Texto simple.
- Sistema. Nombre del sistema que ha obtenido los resultados. Tipo: Texto simple.
- Partición resultados. Partición del conjunto de datos sobre la que se han obtenido los resultados. Tipo: Lista texto.
- Software de Evaluación. Nombre y URL del software utilizado para evaluar los resultados. Tipo: Enlace y URL.

- Publicación. Espacio donde incluir información sobre la publicación donde se han publicado los resultados. Tipo: Texto largo simple.
- Enlace publicación. Enlace a la publicación donde se han publicado los resultados. Tipo: URL.
- Métricas. Listado de métricas disponibles para evaluar el sistema, entre las cuales:

• Precisión	• RMSE	• MLAS	• WAC
• Recall	• Micro precision	• BLEX	• Sentiment Graph
• F1	• Micro recall	• Pearson correla- tion	$F_1 ICM$
• CEM	• Micro F1	• Spearman	• Hierarchical F
• Accuracy	• MAE	• EMR	• Propensity F
• Macro precision	• MAP	• Reliability	• Exact Match
• Macro recall	• UAS	• Sensivity	• MeasureC
• Macro F1	• LAS	• F0.5	• BertScore

3.8. Visualización del contenido

A continuación describimos las páginas del portal que permiten visualizar el contenido del mismo. Las figuras correspondientes se encuentran en el Apéndice B. El contenido de la base de datos del portal se encuentra distribuido en 4 bloques principales, Conjuntos de datos, Tareas, Competiciones y Foros.

Por otro lado, la navegación se ha planificado en 4 áreas de visualización fundamentales:

- Cabecera, donde aparecen un menú principal que permite acceder a las páginas principales del proyecto, el buscador y las opciones de idioma.
- Pie de página, donde se ha reservado un espacio flexible para añadir cuantos elementos de navegación se vea conveniente e incluso añadir elementos dinámicos como formularios si se considera oportuno.
- Contenido, el propio contenido de la web se ha estructurado de tal manera que el usuario pueda navegar a través de los diferentes contenidos y localizar aquella información que le sea útil.
- Buscador, que permite realizar búsquedas directas y que retorna todos los resultados relacionados con la búsqueda, independientemente del tipo de contenido de que se trate: Conjunto de datos, Tarea, Forum, Resultados o Competición.

Más detalles sobre la visualización del contenido se aporta en el Manual de Usuario que se entrega junto con este documento.

3.8.1. Estilo

En cuanto al estilo de la web, tal como se ha comentado anteriormente se ha trabajado siguiendo el Framework Bootstrap, que marca el estilo general de los componentes del proyecto.

Por otro lado, se trabaja con los colores corporativos de la UNED, sede institucional en la que se desarrolla el proyecto.

3.8.2. Portada

A diferencia de los dos portales analizados para el desarrollo de este proyecto, se ha decidido que este portal tenga una página principal o de inicio que introduzca el portal y muestre los elementos más destacados de cada tipo de contenido.

En la Figura 3 se puede observar como se ha definido la portada, en la que tras una breve descripción del proyecto, aparece un espacio para el estado del arte en cifras. Aquí se muestra el número total de tareas, conjuntos de datos, foros y competiciones contenidos en el portal. A continuación se muestra una fila de cajas con las tareas y otra con la ficha de los conjuntos de datos más recientes.

3.8.3. Tareas

La página de tareas (Figura 4 en Apéndice B) muestra las tareas a modo de tarjetas (cards) informativas con los datos básicos que permiten al usuario identificar rápidamente si se trata de una tarea sobre la que pueda querer consultar más información o no. Además, dada la cantidad de elementos de este tipo que se encuentran en la base de datos y para facilitar la búsqueda, se ha añadido una columna de filtros que permiten al usuario seleccionar ciertos foros o tareas.

3.8.4. Conjuntos de datos

En la página de Conjunto de Datos (ver Figura 7) se muestran filas de fichas de los conjuntos de datos. Estos se pueden filtrar por Foro y NLP Topic, permitiendo al usuario acotar los conjuntos de datos que se van a mostrar. El esquema y diseño es similar al de la página de Tareas.

4. Apéndice: Aspectos técnicos relevantes en el desarrollo de proyectos software

En este apéndice se muestran los aspectos técnicos y de documentación relevantes en el proceso de desarrollo de código. Aunque este aspecto queda, en cierto modo, fuera de los ámbitos del convenio, se ha intentado abordar, hasta donde llegan nuestras posibilidades como universidad, todos los puntos sugeridos.

4.1. Mantenibilidad

Durante el proyecto se han seguido los principios básicos de buenas prácticas de programación, adaptadas en cada caso a su lenguaje de programación: Python y PHP. Entre otros, se han tenido en cuenta las convenciones de nombres de variables, clases, formatos de código, etc., así como estilo de comentarios, descripciones de los métodos, etc. Así mismo, se ha creado un repositorio de documentación en GitHub privado donde se ha ido documentando, en su mayor medida, todos los aspectos técnicos de la implementación de las aplicaciones ODESIA: su arquitectura, herramientas utilizadas en el desarrollo con versiones y dependencias, forma de despliegue, etc.

Por último, se ha habilitado un repositorio GitHub privado para el desarrollo de los proyectos de código, lo que permite el trabajo colaborativo de una forma segura y eficaz, a la vez que permite tener copias de seguridad en caso de posibles fallos.

Por último, y aunque todavía no en pleno desarrollo debido a problemas técnicos en nuestra universidad, se han montado dos servidores, uno para desarrollo y otro para producción. Esta diferenciación, que esperamos poder tener lista próximamente nos permitirá evitar errores de estabilidad en las aplicaciones finales, a la vez que agilizar los procesos de despliegue.

4.2. ENS: Esquema Nacional de Seguridad

La UNED, como centro público de enseñanza, está trabajando desde hace tiempo en certificación de su esquema ENS. Tras una reunión con los responsables de ciberseguridad de nuestra universidad, nos indicaron que actualmente existe un borrador y se está a la espera de próxima aprobación. Por otro lado, los responsables nos indicaron que la infraestructura técnica de seguridad en la UNED supera los estándares y requisitos del esquema ENS. En este contexto, y dado que el desarrollo de este proyecto y el despliegue de las aplicaciones se realiza dentro de la infraestructura de la UNED, todas las aplicaciones ODESIA se encuentran sobre el paraguas de seguridad desarrollado por el servicio de ciberseguridad de la UNED.

Además, se han solicitado los certificados de seguridad, que penderán del certificado de seguridad raíz de la UNED, para poder incluir el protocolo https en las aplicaciones, y que esperamos tener próximamente. No obstante, todas las aplicaciones se ejecutan en una red virtualizada privada, a la que solo tienen acceso los contenedores que así se hayan configurados, con un único contenedor con acceso a Internet. Además, este último contenedor se encuentra protegido por el firewall UNED que da cobertura y seguridad a todos los servidores de la Universidad.

4.3. ENI: Esquema Nacional de Interoperabilidad

La misión y objetivos del Esquema Nacional de Interoperabilidad queda un poco fuera del alcance de este proyecto, dadas las dimensiones del mismo y publico objetivo. No obstante, todas las aplicaciones e infraestructuras se están desarrollando teniendo en cuenta que será usadas en entornos de escritorio y con el navegador más popular: Chrome. Dicho esto, se está haciendo un esfuerzo, dentro de nuestras posibilidades, en adaptar todo lo posible las aplicaciones a otros entornos como móviles, tabletas, etc. Así mismo, se ha utilizado la especificación independiente Swagger para la comunicación entre elementos del proyecto ODESIA para conseguir un mejor control y mantenimiento de los mismos.

4.4. Reglamento General de Protección de Datos

Al igual que en el ENS, el proyecto ODESIA se encuentra bajo el amparo del esquema de protección de datos de la UNED.

4.5. Informe de técnicas de Search Engine Optimization

Las aplicaciones ODESIA son aplicaciones cuyo objetivo no es el posicionamiento en los buscadores, su uso por expertos cualificados, por lo que este punto no ha sido tenido en cuenta dentro del proyecto. No obstante, durante el desarrollo web de las aplicaciones se han usado los estándares de HTML5 con las mejores práctica de desarrollo web dentro de los lenguajes de programación y frameworks.

4.6. Diseño de la navegabilidad

El desarrollo de las aplicaciones ODESIA se han centrado desde sus inicios en el desarrollo de aplicaciones amigables y fáciles de usar. Es por ello que la navegabilidad de las mismas está entre sus objetivos principales, proporcionando todas su funcionalidad mediante como mucho tres clicks.

5. Trabajo Futuro

En este documento hemos descrito la Versión 2 del portal ODESIA, que continuará desarrollándose durante el próximo año. En concreto, para el Año 3 del proyecto proponemos las siguientes acciones:

- Poblar la base de datos con datos del 2024.
- Realizar pruebas de uso con diferentes tipos de usuarios para optimizar el acceso a la información del portal en función de diferentes necesidades.
- Mejorar las funcionalidades de introducción, modificación y eliminación de contenidos en el portal, de tal manera que el mantenimiento del mismo pueda ser realizado por personas con conocimientos básicos de informática y sin necesidad de un entrenamiento específico.

6. Conclusiones

En este informe hemos presentado la Versión 2 del portal ODESIA, portal informativo del estado del arte del procesamiento del lenguaje natural en español, desarrollado en el marco del proyecto del Espacio de Observación de Inteligencia Artificial en Español, en concreto Ámbito 1 Estado del arte comparado, Actividad 1.2 Portal del estado del arte en español. El portal web permite acceder a información sobre el estado del arte para las tareas de PLN que se han organizado en foros nacionales e internacionales desde 2013 a 2022 con conjuntos de datos en español. En concreto permite acceder a información sobre conjuntos de datos existentes, competiciones, tareas de PLN y resultados de sistemas. El portal ha sido creado teniendo en cuenta su uso potencial por diferentes tipos de usuarios: investigadores, empresas, instituciones y ciudadanos en general.

La Versión 2 consituye ya una versión completamente funcional del portal, y en la próxima anualidad, el trabajo a realizar estará dirigido a facilitar el mantenimiento del mismo y su actualización.

Agradecimientos

Este trabajo ha sido financiado por la Unión Europea - NextGenerationEU a través del “Plan de Recuperación, Transformación y Resiliencia”, por el Ministerio de Asuntos Económicos y Transformación Digital y por la UNED. Sin embargo, los puntos de vista y las opiniones expresadas son únicamente los del autor o autores y no reflejan necesariamente los de la Unión Europea o la Comisión Europea. Ni la Unión Europea ni la Comisión Europea pueden ser consideradas responsables de los mismos.

Bibliografía

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#).

David Schlangen. 2021. [Targeting the benchmark: On methodology in current natural language processing research](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

A. Apéndice: Modelo de base de datos.

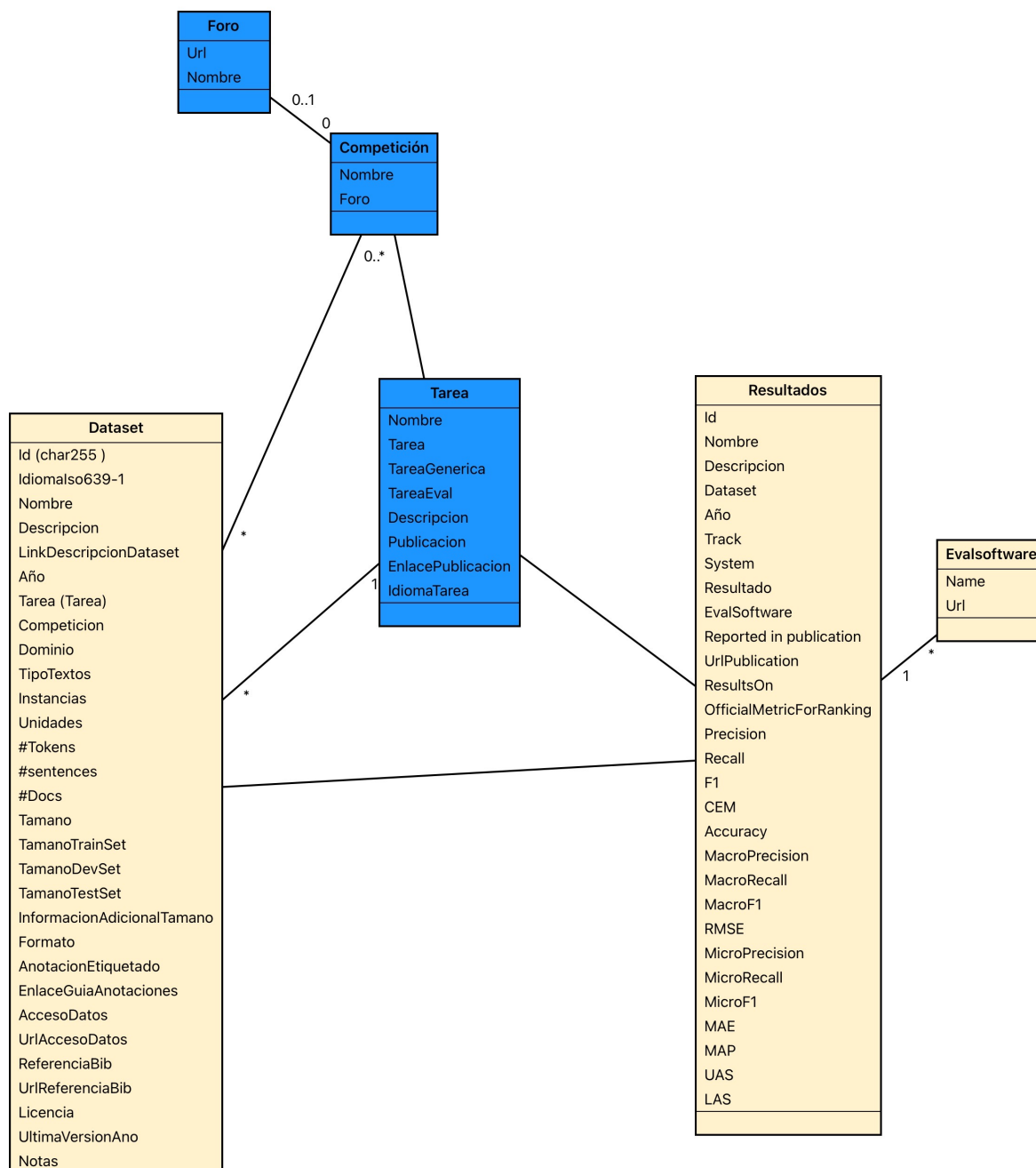


Figura 2: Modelo de la base de datos implementada en Drupal.

B. Apéndice: Páginas web del portal.



Figura 3: Portada del portal ODESIA.



TAREAS

Una tarea es una actividad propuesta con la finalidad de resolver un problema concreto de PLN, generalmente en el marco de una competición. A continuación se muestra información sobre tareas de PLN en Español desde 2013 hasta la actualidad.

Filtrar por

Tarea abstracta

Agrupamiento (2) Clasificación (146)
Correlación (5) Ranking (3)
Regresión (8)

NLP topic

reconocimiento de entidades nombradas (14)
análisis de sentimiento (20)
análisis estilístico (3)
análisis sintáctico (4)
categorización de textos (1)

Foro

CLEF (14) COLING (1) CoNLL (4)
IberEVAL (10) IberLEF (47)
PAN (2) Second CWI Shared Task (1)
SEMEVAL (15)

Dominio

COVID (1) Diversos (1)
Educación (2) Ficción (1)
Finanzas (3) Gastronomía (1)
General (5) Salud (28) Legal (2)
Noticias (8) Social (8) Turismo (6)
Política (6) otros (1)

Año

+ Todos - 2023 2022 2021
2020 2019 2018 2017 2016
2015 <2015

Idioma

Español (153)
Español (Argentina) (2)
Español (Costa Rica) (2)
Español (Cuba) (1)
Español (México) (14)
Español (Peru) (2)

HOMO-MEX: FINE-GRAINED HATE SPEECH DETECTION TRACK
IBERLEF 2023

NLP topic: detección de odio
Dataset: HOMO-MEX 2023
Foro: IberLEF
Competición: HOMO-MEX: Hate speech detection in Online Messages directed tOWards the MEXican spanish speaking LGBTQ+ population
Dominio: Social
Idioma(s): Español, Español (Mexico)

[Ver resultados >>](#)

REST-MEX: THEMATIC UNSUPERVISED CLASSIFICATION
IBERLEF 2023

NLP topic: modelado de temas
Dataset: Rest-Mex 2023 Clustering
Foro: IberLEF
Competición: Rest-Mex 2023: Research on Sentiment Analysis Task for Mexican Tourist Texts
Dominio: Noticias
Idioma(s): Español

[Ver resultados >>](#)

MEDPROCNER: CLINICAL PROCEDURE NORMALIZATION
CLEF 2023

NLP topic: enlace de entidades
Dataset: MedProcNER/ProcTEMIST corpus 2023
Foro: CLEF
Competición: BioASQ 2023: Large-scale Biomedical Semantic Indexing and Question Answering
Dominio: Salud
Idioma(s): Español

[Ver resultados >>](#)

EXIST: SOURCE INTENTION
CLEF 2023

NLP topic: detección de odio
Dataset: EXIST 2023 ES
Foro: CLEF
Competición: EXIST: sEXism Identification in Social neTworks
Dominio: Social
Idioma(s): Español

[Ver resultados >>](#)

MEDPROCNER: CLINICAL PROCEDURE-BASED DOCUMENT INDEXING
CLEF 2023

NLP topic: categorización de textos
Dataset: MedProcNER/ProcTEMIST corpus 2023
Foro: CLEF
Competición: BioASQ 2023: Large-scale Biomedical Semantic Indexing and Question Answering
Dominio: Salud
Idioma(s): Español

[Ver resultados >>](#)

MENTALRISKES - NON-DEFINED DISORDER DETECTION - SIMPLE...
IBERLEF 2023

NLP topic: profiling
Dataset: MentalRiskES - Undefined disorder 2023
Foro: IberLEF
Competición: MentalRiskES: Early detection of mental disorders risk in Spanish
Dominio: Salud
Idioma(s): Español

[Ver resultados >>](#)

HUHU - DEGREE OF PREJUDICE PREDICTION
IBERLEF 2023

NLP topic: procesamiento de humor
Dataset: HUHU 2023
Foro: IberLEF
Competición: HUHU: Hurtful HUmour - Detection of humour spreading prejudice on Twitter
Dominio: Social
Idioma(s): Español

[Ver resultados >>](#)

MENTALRISKES - DEPRESSION DETECTION - SIMPLE REGRESSION
IBERLEF 2023

NLP topic: profiling
Dataset: MentalRiskES - Depression 2023
Foro: IberLEF
Competición: MentalRiskES: Early detection of mental disorders risk in Spanish
Dominio: Salud
Idioma(s): Español

[Ver resultados >>](#)

DA-VINCIS - DETECTION OF AGGRESSIVE AND VIOLENT INCIDENTS FR...
IBERLEF 2023

NLP topic: procesamiento de eventos
Dataset: DA-VINCIS 2023
Foro: IberLEF
Competición: DA-VINCIS: Multimodal Information for the Detection of Aggressive and Violent INCidents from Social media in Spanish
Dominio: Social
Idioma(s): Español, Español (Mexico)

[Ver resultados >>](#)

JOKER: PUN LOCATION IN SPANISH
CLEF 2023

NLP topic: procesamiento de humor
Dataset: JOKER 2023 ES
Foro: CLEF
Competición: JOKER: Automatic Wordplay Analysis
Dominio: Social
Idioma(s): Español

[Ver resultados >>](#)

FINANCES: FINANCIAL TARGETED SENTIMENT ANALYSIS
IBERLEF 2023

NLP topic: análisis de sentimiento
Dataset: FinancES 2023
Foro: IberLEF
Competición: FinancES: Financial Targeted Sentiment Analysis in Spanish
Dominio: Finanzas
Idioma(s): Español

[Ver resultados >>](#)

CLINAIIS: AUTOMATIC IDENTIFICATION OF SECTIONS IN CLINICAL...
IBERLEF 2023

NLP topic: clasificación de textos
Dataset: ClinAIS 2023
Foro: IberLEF
Competición: ClinAIS: Automatic identification of sections in clinical documents
Dominio: Salud
Idioma(s): Español

[Ver resultados >>](#)

Figura 4: Página 'Tareas' del portal ODESIA.

Inicio
Tareas
Datasets
Competiciones
Foros
Buscar
Idioma

DATASETS

A continuación se muestra información sobre conjuntos de datos textuales en español creados con el objetivo de resolver tareas de PLN. En este caso, se trata de colecciones de textos, generalmente enriquecidas con anotaciones.

Filtrar por

Domínio

COVID (1) Diversos (1)

Educación (2) Ficción (1)

Finanzas (3) Gastronomía (1)

General (5) Salud (28) Legal (2)

Noticias (8) Social (8) Turismo (6)

Política (6) otros (1)

NLP topic

reconocimiento de entidades nombradas (14)

análisis de sentimiento (20)

análisis estilístico (3)

análisis sintáctico (4)

categorización de textos (1)

Idioma

Español (153)

Español (Argentina) (2)

Español (Costa Rica) (2)

Español (Cuba) (1)

Español (Mexico) (14)

Español (Peru) (2)

Año

- Todos - 2023 2022 2021

2020 2019 2018 2017 2016

2015 <2015

JOKER 2023 ES

Social

🌐 Español , 🌐 Inglés

📅 Publicado en 2023

📄 4,235

📄 Tuits

procesamiento de humor

JOKER 2023 ES

Social

🌐 Español , 🌐 Inglés

📅 Publicado en 2023

📄 4,235

📄 Tuits

procesamiento de humor

EXIST 2023 ES

Social

🌐 Español , 🌐 Inglés

📅 Publicado en 2023

📄 2,299

📄 Tuits

detección de odio

EXIST 2023 ES

Social

🌐 Español , 🌐 Inglés

📅 Publicado en 2023

📄 2,299

📄 Tuits

detección de odio

CT-CWT-23-ES

Noticias

🌐 Español

📅 Publicado en 2023

📄 29,984

📄 Tuits

detección de noticias falsas

MedProcNER/ProcTEMIST corpus 2023

Salud

🌐 Español

📅 Publicado en 2023

📄 1,000

📄 Informes clínicos

reconocimiento de entidades nombradas

MedProcNER/ProcTEMIST corpus 2023

Salud

🌐 Español

📅 Publicado en 2023

📄 1,000

📄 Informes clínicos

enlace de entidades

MedProcNER/ProcTEMIST corpus 2023

Salud

🌐 Español

📅 Publicado en 2023

📄 1,000

📄 Informes clínicos

categorización de textos

Rest-Mex 2023 Clustering

Noticias

🌐 Español , 🌐 Español (Mexico)

📅 Publicado en 2023

📄 114,550

📄 Noticias

modelado de temas

Rest-Mex 2023 Sentiment

Turismo

🌐 Español (Mexico)

📅 Publicado en 2023

📄 359,565

📄 Opiniones Tripadvisor

análisis de sentimiento

FinancES 2023

Finanzas

🌐 Español

📅 Publicado en 2023

📄 7,980

📄 Titulares de noticias

análisis de sentimiento

FinancES 2023

Finanzas

🌐 Español

📅 Publicado en 2023

📄 7,980

📄 Titulares de noticias

análisis de sentimiento

PoliticES 2023

Social, Política

🌐 Español

📅 Publicado en 2023

📄 2,797

📄 Tuits

profiling

HUHU 2023

Social

🌐 Español

📅 Publicado en 2023

📄 3,449

📄 Tuits

procesamiento de humor

1 2 3 4 5 6 7 8 9 ... » Último »

Figura 5: Página ‘Datasets’ del portal ODESIA.



COMPETICIONES

Filtrar por

Forum

CLEF (14) COLING (1) CoNLL (4)
IberEVAL (10) IberLEF (47)
PAN (2) Second CWI Shared Task (1)
SEMEVAL (15)

Year

- Todos - 2023 2022 2021
2020 2019 2018 2017 2016
2015 <2015

AuTexTification: Automated Text Identification

IBERLEF

Año: 2023

<https://sites.google.com/view...>

Ver más >>

BioASQ 2023: Large-scale Biomedical Semantic Indexing and Question Answering

CLEF

Año: 2023

<http://www.bioasq.org/worksho...>

Ver más >>

CheckThat!: Check-Worthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and their Sources

CLEF

Año: 2023

<http://checkthat.gitlab.io/>

Ver más >>

ClinAIS: Automatic identification of sections in clinical documents

IBERLEF

Año: 2023

<http://ixa2.si.ehu.eus/clinai...>

Ver más >>

DA-VINCIS: Multimodal Information for the Detection of Aggressive and Violent INCidents from Social media in Spanish

IBERLEF

Año: 2023

<https://sites.google.com/view...>

Ver más >>

DIPROMATS: Automatic Detection Of Propaganda TechniquEs from Diplomats in Spanish

IBERLEF

Año: 2023

<https://sites.google.com/view...>

Ver más >>

Figura 6: Página ‘Competiciones’ del portal ODESIA.



FOROS

CLEF

Conference and Labs of the Evaluation Forum
<https://www.clef-initiative.eu/>

El objetivo del CLEF es fomentar la investigación, la innovación y el desarrollo de sistemas de acceso a la información, haciendo hincapié en la información multilingüe y multimodal con distintos niveles de estructura.

[más información >>](#)

IberLEF es una campaña de evaluación que pretende fomentar la definición de nuevos retos en la comunidad investigadora de Procesamiento del Lenguaje Natural y la obtención de resultados punteros, involucrando al menos una de las lenguas ibéricas: Español, Portugués, Catalán, Euskera o Gallego. Por ello, bajo el paraguas de IberLEF se organizan anualmente varias competiciones, cuyos resultados se presentan en el congreso anual de la Sociedad Española de Procesamiento del Lenguaje Natural (SEPLN).

[más información >>](#)

IBERLEF

Evaluation campaign for Natural Language Processing (NLP) systems in Spanish and other Iberian languages
<https://sites.google.com/view/iberlef-2023/home>

IberEval

Workshop on Evaluation of Human Language Technologies for Iberian Languages

IberEval tenía como objetivo fomentar y promover el desarrollo de las Tecnologías del Lenguaje Humano (HLT) para las lenguas ibéricas (español, portugués, catalán, euskera y gallego), mediante la creación de una serie de evaluaciones y un foro de debate sobre sistemas de comunicación. Actualmente ha sido sustituido por IberLEF.

[más información >>](#)

SemEval es una serie de talleres internacionales de investigación sobre procesamiento del lenguaje natural (PLN), cuya misión es avanzar en el estado del arte del análisis semántico y ayudar a crear conjuntos de datos anotados de alta calidad sobre una variedad de problemas cada vez más desafiantes. Cada año, SemEval acoge un conjunto de competencias en las que se presentan y comparan sistemas de análisis semántico computacional diseñados por diferentes equipos.

[más información >>](#)

SEMEVAL

International Workshop on Semantic Evaluation
<https://semeval.github.io/>

Figura 7: Página 'Foros' del portal ODESIA.