Plan Overview

A Data Management Plan created using DMP Tool

Title: Clasificador entre lenguaje técnico médico y PLS

Creator: Carlos Urrego

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Julian Rico, Cristian Romero, Luis Ángel Sánchez Marín

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

En el ámbito de la comunicación médica, existe una brecha significativa entre el lenguaje técnico utilizado por profesionales de la salud y los resúmenes en lenguaje sencillo (Plain Language Summary, PLS) dirigidos a pacientes y público general. Este proyecto propone explorar,a modo de prueba de concepto, el desarrollo de un clasificador automático de texto basado en aprendizaje automático para distinguir entre estos dos tipos de lenguaje, facilitando la accesibilidad y comprensión de información médica.

Start date: 08-11-2025

Last modified: 08-18-2025

Clasificador entre lenguaje técnico médico y PLS

Data Collection

What data will you collect or create?

Para el proyecto, hemos identificado un repositorio inicial del Center for Open Science, una ONG dedicada a la ciencia abierta. El dataset contiene pares de textos que comparan jerga médica con sus adaptaciones a lenguaje simple, todas redactadas por humanos.

Este conjunto de datos se usará para entrenar un modelo de aprendizaje supervisado, utilizando el texto complejo y su versión simplificada de forma etiquetada.

How will the data be collected or created?

El Dataset actual verificado está disponible aquí: https://osf.io/rnpmf/files/osfstorage

Los datos de entrenamiento equivalen a 636 registros, por lo que, se usará para un test inicial, si el modelo no logra pasar el rendimiento esperado se considerará el uso de otras fuentes para la ampliación del dataset como la colaboración Cochrane que es una organización que realiza revisiones sistemáticas de investigación médica y crea resúmenes en lenguaje sencillo para pacientes

Documentation and Metadata

What documentation and metadata will accompany the data?

La estrategia de documentación y gestión de metadatos se adaptará a la evolución del conjunto de datos. En la fase actual, que emplea una única fuente, se mantendrá la metainformación proporcionada por el repositorio original para preservar su integridad.

Para futuras fases que contemplen la integración de datos de múltiples fuentes, se implementará un protocolo de documentación riguroso. Este protocolo garantizará la trazabilidad y reproducibilidad mediante la creación de los siguientes artefactos: un archivo README.md descriptivo, un diccionario de datos exhaustivo y un archivo LICENSE.txt. Este último gestionará las licencias de las fuentes de manera compatible, atribuyendo siempre el crédito a los autores originales y especificando metadatos técnicos como el formato de archivo y la versión.

Ethics and Legal Compliance

How will you manage any ethical issues?

Para gestionar los aspectos éticos, adoptaremos un enfoque proactivo. Fundamentalmente, garantizaremos la privacidad asegurando que todos los conjuntos de datos estén completamente anonimizados, eliminando cualquier Información de Identificación Personal (PII) como nombres o identificadores de pacientes. Si bien los datos estarán libres de PII, reconocemos que el contenido médico (diagnósticos, síntomas, etc.) sigue siendo de naturaleza sensible. Por lo tanto, todo el dataset será tratado con precuación y el modelo resultante se diseñará como una herramienta de apoyo informativo y nunca como un sustituto del consejo médico profesional, para prevenir cualquier daño derivado de malas interpretaciones.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

La gestión de los derechos de propiedad intelectual y copyright se realizará de forma rigurosa y ética, utilizando exclusivamente conjuntos de datos y software cuyas licencias (como Creative Commons, MIT o Apache 2.0) permitan explícitamente su uso para la investigación y la creación de obras derivadas. Se documentará y respetará la licencia de cada fuente, garantizando siempre la correcta atribución a los autores originales en toda la documentación del proyecto.

En caso de combinar datos de múltiples procedencias, se llevará a cabo un análisis de compatibilidad de licencias para asegurar el cumplimiento de todos los términos.

Storage and Backup

How will the data be stored and backed up during the research?

Los datos del proyecto se almacenarán principalmente en la plataforma OSF, que proporciona un entorno seguro y gratuito, durante la investigación se implementará una estrategia de respaldo orientada a mantener al menos tres copias de los datos, en la práctica, los respaldos se generarán de forma automática y periódica, evitando depender únicamente de dispositivos personales como laptops o discos duros, que representan un alto riesgo de pérdida o corrupción, con este enfoque se garantiza la integridad, accesibilidad y preservación de los datos a lo largo de toda la investigación.

How will you manage access and security?

El acceso se controlará a través de los permisos de OSF, que permiten establecer distintos niveles de visibilidad (público, restringido a colaboradores o privado), para esta investigación los datos permanecerán en modo privado, de forma que solo el equipo de investigación podrá acceder a ellos.

Además, se evitará el almacenamiento de datos sensibles en servicios de terceros que no cumplan con las políticas y normativas de protección de datos, en el caso de copias adicionales, la información se cifrará, especialmente cuando se almacene en servicios externos o dispositivos físicos, con el fin de garantizar la confidencialidad y seguridad.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

El dataset principal, que constituye el primer corpus alineado a nivel de documento y oración, se conservará de manera indefinida por su valor científico y potencial de reutilización. También se preservarán los metadatos asociados (identificadores de preguntas, categorías clínicas o biológicas, pmids y anotaciones de los adaptadores), ya que resultan esenciales para la interpretación y reutilización del dataset en nuevos contextos, la documentación complementaria se mantendrá junto al dataset con el fin de garantizar la reproducibilidad, la transparencia y la correcta preservación a largo plazo.

What is the long-term preservation plan for the dataset?

El dataset se preservará en el repositorio OSF, que ofrece un entorno seguro y sostenible para el almacenamiento a largo plazo sin costo. Adicionalmente, se considerará el uso de soluciones institucionales de almacenamiento como medida secundaria para garantizar la redundancia de los datos, todos los archivos con valor científico a largo plazo (el corpus alineado, los metadatos y la documentación) se prepararán para su preservación en formatos abiertos, con el fin de maximizar su accesibilidad y reutilización

Data Sharing

How will you share the data?

Los datos estarán disponibles de forma abierta y sin restricciones de uso para cualquier interesado, incluyendo investigadores, desarrolladores y profesionales del ámbito médico.

Los datos se compartirán a través de un repositorio público en la herramienta DMPTools, que permite una gestión estructurada y accesible de conjuntos de datos. Los usuarios podrán descargar los datos directamente desde el repositorio, acompañado de documentación que detalle su estructura, recolección y uso.

Are any restrictions on data sharing required?

Dado que los datos utilizados en el proyecto podrían incluir historiales médicos o información de identificación personal (PII), se implementará un proceso riguroso de anonimización para eliminar cualquier dato sensible antes de su publicación. Este proceso garantizará que los textos resultantes puedan compartirse de manera abierta y sin restricciones, cumpliendo con las normativas de privacidad y protección de datos.

Además, considerando que los datos podrían provenir de estudios previos, la base de datos final heredará las restricciones de uso de las fuentes originales. Por ello, se seleccionarán exclusivamente fuentes que cuenten con acuerdos de uso que permitan la compartición abierta y sin limitaciones, asegurando que los datos derivados puedan ser utilizados libremente por la comunidad académica, profesional y el público en general.

Responsibilities and Resources

Who will be responsible for data management?

Los 4 miembros del equipo se comprometen a seguir y velar por la implementación del DMP, asi como a reportar al resto del equipo cualquier desviación que se pueda surgir. Las decisiones significativas acerca de cambios en el mismo que se puedan presentar o necesitar se harán tras discutir el asunto en equipo.

Debido a la elevada carga académica sumada a las responsabilidades ajenas a la academia de cada integrante del equipo, no se asignan de antemano responsabilidades ligadas a tareas especificas. En su lugar nos adherimos al cronograma de tareas propuesto en la descripción del proyecto con revisiones semanales y asignación adhoc de paquetes de trabajo.

What resources will you require to deliver your plan?

El manejo de los datos se implementará mediante repositorios y versionado siguiendo principios de MLOps. Teniendo en cuenta la naturaleza de los datos que se van a utilizar (tipo y tamaño), los datos utilizados para el entrenamiento y la evaluación de los modelos, en todas sus etapas, se almacenarán y versionarán con una herramienta de control de datos de uso libre (p. ej., DVC). Esto implica una curva de aprendizaje para su uso correcto y reproducible.

Los modelos y experimentos se gestionarán con MLflow (ejecuciones, parámetros, métricas y artefactos), lo que también requiere un proceso de adopción y configuración.

La documentación de soporte y el código generado se alojarán en un repositorio Git en el cual solamente los 4 integrantes participarán como colaboradores.

Planned Research Outputs

Model representation - "Determinación del modelo de clasificacion de textos PLS"

Explorar modelos de clasificación de texto apoyandonos en la literatura bibliográfica para asi contrastar defierentes clasificadores y obtener uno capaz de distinguir entre lenguaje técnico y lenguaje sencillo en el contexto médico.

Workflow - "Exploración de técnicas de procesamiento de texto"

Explorar diferentes aproximaciónes a la tarea de preparar (preprocesamiento) datos de naturaleza textual con el objetivo de clasificar textos técnicos vs no tecnicos en el contexto meédico.

Software - "Despliegue del clasificador de manera nativa en la nube"

Experimentar de primera mano el proceso llevar un prototipo funcional a un ambiente de producción, el cual puede ser consumido por cualquier usuario, usando principios de desarrollo de software y soluciones de inteligencia artificial como MLOps.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Determinación del modelo de clasificacion de texto ...	Model representation	Unspecified	Open	None specified		None specified	None specified	Yes	No
Exploración de técnicas de procesamiento de texto	Workflow	Unspecified	Open	None specified		None specified	None specified	No	No
Despliegue del clasificador de manera nativa en la ...	Software	Unspecified	Open	None specified		None specified	None specified	Yes	No