Análisis y Automatización del Procesamiento de Requerimientos de la Ley de Transparencia en el Ministerio de Salud utilizando Técnicas de Procesamiento del Lenguaje Natural (NLP)

Estudiante:	Luis Figueroa G.
Profesor Guía:	Patricio Wolff R.
Profesor Co Guía:	Sebastián Ríos
Fecha:	26 de abril de 2024

1. Antecedentes

La Ley 20.285 sobre el Acceso a la Información Pública, más conocida como Ley de Transparencia, establece el derecho fundamental de cualquier persona a solicitar y obtener información de los órganos del Estado[1]. En el caso del Ministerio de Salud (Minsal), este derecho se traduce en la posibilidad de acceder a una amplia gama de datos e información relacionada con su funcionamiento y las políticas públicas en materia de salud.

En este contexto, el Minsal tiene la obligación de procesar las solicitudes de información que se le presenten a través del Portal de Transparencia. Para ello, cuenta con una Unidad de Transparencia encargada de revisar y tramitar estos requerimientos. Sin embargo, el proceso actual presenta ciertas ineficiencias debido a la diversidad de tipos de solicitudes, las cuales requieren un tratamiento diferenciado. Entre ellas encontramos:

- **Solicitudes realizadas de forma incorrecta en el portal**, lo que genera retrasos en la atención de aquellas solicitudes válidas.
- **Solicitudes que demandan un trabajo de búsqueda y recopilación de datos considerable,** lo que prolonga los tiempos de respuesta.
- **Solicitudes de información que ya se encuentra publicada en el portal o en otras plataformas de acceso público,** lo que genera un uso ineficiente de los recursos de la Unidad de Transparencia.

En los últimos años, el uso de técnicas de NLP ha experimentado un crecimiento exponencial, impulsado por su implementación como soluciones y servicios en diversas empresas. El mercado global de NLP alcanzó un tamaño de USD 14.53 mil millones en 2022, y se proyecta que alcance los USD 131.33 mil millones para el año 2031[2]. Esta tendencia y el creciente desarrollo de este tipo de tecnologías, presenta una oportunidad invaluable para el Ministerio de Salud (Minsal), ya que permite aprovechar el poder del NLP para optimizar el proceso de atención de solicitudes de información a través del Portal de Transparencia.

2. Objetivos

2.1. Objetivo General

Disminuir los tiempos de procesamiento de las solicitudes recibidas por el portal de transparencia del Ministerio de Salud mediante el desarrollo de una herramienta automatizada que incorpore técnicas de NLP para mejorar la eficiencia en la gestión de los recursos ministeriales

2.2. Objetivos Específicos

- **Preparación de datos:** Analizar y limpiar datos para asegurar que la información esté libre de errores y sea relevante para el análisis posterior, garantizando la privacidad de los datos sensibles.
- **Entrenamiento de modelos NLP:** Entrenar modelos de Procesamiento del Lenguaje Natural (NLP) para identificar y clasificar las diferentes solicitudes recibidas, incluyendo la selección de algoritmos y el ajuste de hiperparámetros para maximizar la precisión del modelo.
- **Evaluación de resultados:** Analizar los resultados obtenidos mediante métricas específicas y relevantes para el problema planteado. Además, se llevará a cabo una comparación con modelos del Estado del Arte para contextualizar los hallazgos.
- **Análisis y optimización:** Realizar análisis detallados de los resultados obtenidos, con el fin de encontrar oportunidades de mejora con respecto a la disponibilización de información y administración de recursos.

3. Metodologías

El desarrollo de este proyecto se basa en la metodología CRISP-DM, un enfoque iterativo para la construcción de modelos de machine learning. La preparación de los datos, crucial para el éxito del proyecto, exige un enfoque centrado en la privacidad, implementando medidas de seguridad y anonimización adecuadas. Se utilizarán modelos de redes neuronales, que son menos sensibles a la calidad de los datos, sin embargo, es fundamental realizar un buen preprocesamiento de los datos para reducir el ruido de estos, además de un etiquetado de alta calidad para generar un conjunto de datos confiable y preciso.

Con respecto a modelamiento, se evaluarán dos enfoques distintos:

- **Named Entity Recognition (NER):** El NER es una técnica de NLP que permite identificar nombres de entidades en el texto que se asemejen a categorías predefinidas, como nombres de personas, ubicaciones y organizaciones[3]. En este contexto permitiría identificar entidades como personas, hospitales, tipos de datos, entre otros.
- **Multilabel Classification:** En entornos médicos y de salud, se ha demostrado que el fine-tuning de modelos de redes neuronales preentrenados para tareas de clasificación multiclase produce resultados satisfactorios. Un ejemplo destacado de esto es el modelo PubMedBERT[4]. Este tipo de implementación permitirá clasificar de manera automática los diferentes tipos de solicitudes que recibe el Minsal.

4. Resultados Esperados

Se espera de este proyecto, crear un conjunto de datos de alta calidad que garantice la privacidad de los usuarios del Portal de Transparencia. Asimismo, se espera desarrollar una herramienta automatizada basada en modelos de Procesamiento del Lenguaje Natural robustos y con un buen rendimiento, evaluado mediante métricas pertinentes. Además, esta herramienta, y especialmente los modelos entrenados, deben requerir recursos computacionales que se ajusten a las capacidades disponibles en el Minsal, permitiendo su implementación eficiente en un entorno de producción, de manera que la herramienta contribuya a reducir los tiempos de respuesta y a generar información relevante para el caso en cuestión.

5. Plan de Trabajo

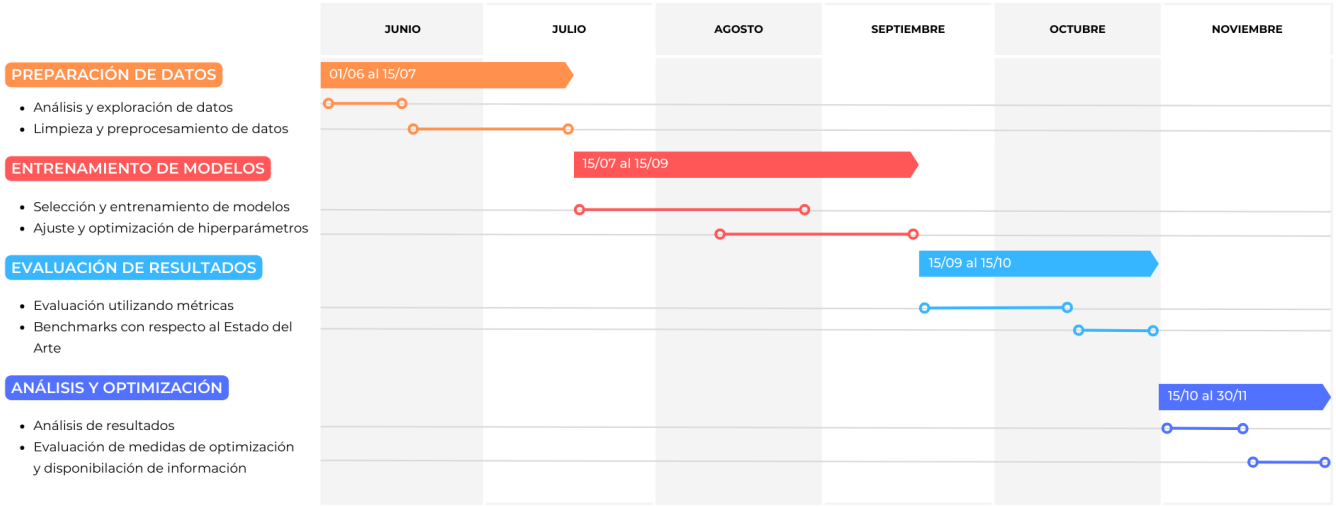

Figura 1: Carta Gantt asociada al plan de trabajo para el desarrollo de la tesis.

6. Resultados Preliminares

Se ha realizado un análisis preliminar de aproximadamente 8.000 solicitudes ingresadas al Portal de Transparencia del Ministerio de Salud entre 2015 y 2024. Este conjunto de datos presenta una gran diversidad, desde consultas específicas como la cantidad de metros cuadrados construidos en hospitales a nivel nacional hasta solicitudes sobre el nombre del Ministro de Salud. La diversidad de las solicitudes exige un enfoque flexible y adaptable para su análisis, mientras que la privacidad de los datos personales es una prioridad fundamental. Por ello, se implementarán medidas de seguridad y anonimización adecuadas durante el proceso de análisis.

Referencias

[1] **República de Chile.** (2008). *Ley 20.285 Sobre el Acceso a la Información Pública. Diario Oficial de la República de Chile, p. 1.*
https://www.leychile.cl/navegar?idNorma=276363

[2] **Analytics India Magazine.** (2022). *Global Natural Language Processing (NLP) Market Research Report*
https://www.insightaceanalytic.com/report/natural-language-processing-nlp-market/2086

[3] **Zhang, Y., & Xiao, G.** (2024). *Named Entity Recognition Datasets: a Classification Framework, The International Journal of Computational Intelligence Systems, p. 1.*
https://doi.org/10.1007/s44196-024-00456-1

[4] **Chen, Qingyu & Du, Jingcheng & Hu, Yan & Keloth, Vipina & Peng, Xueqing & Raja, Kalpana & Zhang, Rui & Lu, Zhiyong & Qi, Wang.** (2023). *Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. p. 8.*
https://www.researchgate.net/publication/371123686_Large_language_models_in_biomedical_natural_language_processing_benchmarks_baselines_and_recommendations

Firma Estudiante	**Firma Profesor Guía**	**Firma Profesor Co-Guía**
Luis Figueroa G.	Patricio Wolff R.	Sebastián Ríos