**Aplicación de herramientas de inteligencia artificial en el análisis de las normas regulatorias elaboradas por la Autoridad Regulatoria Nuclear**

**A. Gómez\***(1), **V. Bettachini**(2), **H. Costa**(2), **M. Bringas**(2)e **I. Nesterov**(2)

*(1) Subgerencia Normativa Regulatoria, Autoridad Regulatoria Nuclear (ARN), Av. del Libertador 8250, Buenos Aires, Argentina*

*(2) Maestría en explotación de datos y descubrimiento del conocimiento, FCEyN-FI, UBA, Ciudad Universitaria, Buenos Aires, Argentina.*

\*Correo Electrónico (autor de contacto): [agomez@arn.gob.ar](mailto:autor@mmm.gob.ar)

**Tópico:** Normativa

Implementar herramientas de inteligencia artificial (IA), especialmente aquellas fundamentadas en el procesamiento del lenguaje natural (NLP), en un ámbito técnico específico cono las normas regulatorias nucleares, requiere expandir el alcance de los modelos entrenados con corpus de texto general. Este trabajo, desarrollado en el marco de la Subgerencia de Normativa Regulatoria de la Autoridad Regulatoria Nuclear (ARN), organizó la normativa vigente como un grafo dirigido de citas y exploró un método para hacerla más accesible mediante la técnica de generación aumentada por recuperación (RAG), apoyada en grandes modelos de lenguaje (LLM) de acceso libre.

El NLP es un subcampo clave de la inteligencia artificial (IA). Una de sus herramientas más extendidas, los LLM como ChatGPT o Gemini, se entrenan con grandes corpus de datos, como el conjunto de Wikipedia. En tales corpus la terminología técnica es escasa, por lo que para dominios especializados, como la actividad nuclear, los LLM suelen ser insuficientemente precisos [1].

En este trabajo presentamos dos aplicaciones que demostraron resultados en el tratamiento de las 64 normas y 10 guías regulatorias publicadas por la ARN [2]. En conjunto, estas regulaciones rigen la actividad nuclear en la República Argentina. Para construir el corpus de texto, es decir, el conjunto de datos requerido para el NLP, se realizó un web scrapping del sitio de ARN que publica los documentos en formato PDF. Se utilizó la biblioteca Selenium del lenguaje Python para descargarlos secuencialmente y la biblioteca PyPDF2 para extraer su texto plano.

Con el objetivo de estudiar el clustering de los documentos según su temática, por ejemplo, 'Seguridad', 'Protección radiológica', 'Criterios de diseño', etc., se buscó construir un grafo dirigido en el que cada documento actúa como nodo y las citas como aristas. La detección automática de citas a otras normas mediante expresiones regulares, permitió crear un diccionario de citas cruzadas. Utilizando la biblioteca networkx, se generaron representaciones gráficas que facilitaron la cuantificación del grado de agrupamiento. La detección automática de un error topológico en el grafo de citas destacó el potencial del uso de la IA en el tratamiento de la documentación normativa, indicando una falta de actualización de la norma AR 4.8.1 que cita a la norma AR 3.6.1, la cual fue subsumida anteriormente en la norma AR 10.6.1.

Luego del análisis, se construyó un asistente basado en un LLM que responde en lenguaje natural a consultas sobre información contenida en la normativa. Los precursores de los LLM actuales, contaban con, a lo sumo, 3 millones de parámetros, lo que requería un ajuste fino supervisado en el corpus del texto de un dominio particular. Con el advenimiento de los LLM con más de 2 mil millones de parámetros, el reentrenamiento se ha vuelto muy demandante en términos de recursos.

El consenso actual es que si el corpus general de entrenamiento no produce un desempeño adecuado en un dominio específico, es recomendable construir un sistema de recuperación de texto RAG, que añade al LLM la capacidad de interactuar con una base de datos [3]. Este es el caso de la normativa nuclear, cuyo nivel de especificidad la excluye de los corpus de entrenamiento utilizados por los generadores de los LLM con más de 3 mil millones de parámetros, como GPT-4o (OpenAI), Claude (Anthropic), Gemini (Google) o Llama 3.1 (Meta), entre otros. Este último ha sido seleccionado para este trabajo, ya que, hasta la fecha, muestra el mejor desempeño en ensayos públicos con corpus en español [4].

Para este trabajo, se amplió el alcance de este LLM al almacenar en una base de datos la representación vectorial (embeddings) de los tokens, que son los segmentos en que el LLM divide el texto del corpus de normativas. El entorno LangChain para lenguaje Python gestiona esa base, denominada Chroma, así como el recorte en tokens del corpus. Luego, solo es necesario ejecutar el recuperador (retriever) con una plantilla que instruye al modelo a personificar a un experto en normativa nuclear, para habilitar al usuario para realizar consultas (prompts) al RAG relacionadas con la normativa.

**Referencias**

[1] Dima, Alden, Sarah Lukens, Melinda Hodkiewicz, Thurston Sexton, y Michael P. Brundage. 2021. “Adapting Natural Language Processing for Technical Text”. *Applied AI Letters* 2 (3): e33.<https://doi.org/10.1002/ail2.33>.

[2] Autoridad Regulatoria Nuclear. h[ttps://www.argentina.gob.ar/arn/instalaciones-practicas-y-personal-regulado/marco-regulatorio/normas](https://www.argentina.gob.ar/arn/instalaciones-practicas-y-personal-regulado/marco-regulatorio/normas)

[3] Aditya Jain, Amir Maleki, y Nathalie Saade. 2024. “To Fine-Tune or Not to Fine-Tune”. *Meta AI* (blog). 7 de agosto de 2024.<https://ai.meta.com/blog/when-to-fine-tune-llms-vs-other-techniques/>.

[4] “La Leaderboard - a Hugging Face Space by la-leaderboard”, Hugging Face. <https://huggingface.co/spaces/la-leaderboard/la-leaderboard>