

LLM para la Búsqueda de Información Legislativa de la HCDN

Bartolomeo Adrian Gonzalez

Resumen—Este estudio aborda las dificultades en el acceso a la información legislativa en la Honorable Cámara de Diputados de la Nación. Exploramos el potencial de las tecnologías de procesamiento de lenguaje natural, priorizando el método de RAG sobre el Fine-Tuning, debido al carácter de la información a consultar. Se desarrolla y explica el funcionamiento de un sistema detallando cada componente de tal y su objetivo para lograr el correcto funcionamiento de un modelo de lenguaje que logre extraer y otorgar información a los ciudadanos. A pesar de los desafíos presentados, la investigación demuestra el potencial de estas tecnologías para promover la transparencia y la participación ciudadana en el proceso legislativo, facilitar el acceso a la información de carácter público y abrirle las puertas de esa información al mundo.



Figura 1. HCDN

I. INTRODUCCIÓN

La Honorable Cámara de Diputados de la Nación Argentina le ofrece a cualquier individuo con acceso a internet la posibilidad de consultar a través de un sitio web[1] los proyectos de Ley de la Nación Argentina. Este recurso es fundamental para la transparencia gubernamental y la participación ciudadana, permitiendo a los ciudadanos, académicos y profesionales del derecho acceder a información legislativa vital.

Un inconveniente notable, es la dificultad para encontrar información de carácter legislativo en el sitio web, ya sea por lentitud o poca facilidad de uso, esto plantea desafíos significativos para los ciudadanos y los profesionales del derecho en encontrar la información que buscan.

Diversos esfuerzos se han realizado para abordar esta problemática, desde la creación de portales web gubernamentales hasta el desarrollo de herramientas de consulta. Sin embargo, aún persisten obstáculos en términos de usabilidad y accesibilidad de la información.

El gran cambio de tecnologías a través de los años y los cambios de gestión política han llevado a que la información se encuentre descentralizada a través de distintas páginas web.[2][3]

En este contexto, surge la idea de utilizar tecnologías de procesamiento de lenguaje natural[4]. La aplicación de modelos de lenguaje[5] y técnicas de Text-to-SQL[6][7][8] pueden ofrecer una solución para comprender y consultar información legislativos de manera más efectiva y eficiente.

Esta investigación busca abordar esta problemática mediante la implementación de un sistema que integre tecnologías de procesamiento de lenguaje natural[9] y un diseño centrado en el usuario.

Mediante la centralización de los datos, se utilizarán varios componentes posteriormente detallados[10], que en conjunto implementarán un sistema a través del cual cada individuo podrá consultarle a un modelo de lenguaje natural preguntas relacionadas con información legislativa.

Posteriormente se pondrán a prueba los distintos métodos[11][12][13] de obtención de información a través de un modelo de lenguaje natural, así como también se comprobará que modelos de lenguaje[14] se encuentran disponibles actualmente en el mercado y cuáles son las propiedades de cada uno, con el fin de determinar cuál es el procedimiento adecuado para el contexto en el que se encuentra la problemática.

El modelo de lenguaje tendrá la capacidad de comportarse y responder como un asistente legislativo. Se le proveerá a dicho modelo el contexto[15] necesario para responder en el caso que se le sea requerida cierta información.

Las secciones de la investigación se estructuran de la siguiente manera:

- Introducción
- Motivación
- Large Language Model
- Marco teórico
- Componentes
- Resultado
- Conclusión y trabajos futuros

Con esta estructura, se espera proporcionar una comprensión integral de cómo las tecnologías de procesamiento de lenguaje natural pueden transformar el acceso a la información legislativa, facilitando el uso de las herramientas de consulta.

II. MOTIVACIÓN

En la Honorable Cámara de Diputados de la Nación Argentina existe una dificultad para consultar y acceder a la información legislativa de manera eficaz y eficiente a través del sitio web oficial. Esta dificultad puede atribuirse a varios factores, que van desde la lentitud del sitio hasta la dificultad de uso de las herramientas de búsqueda y navegación disponibles.

La importancia de abordar este problema afecta múltiples sectores de la sociedad. Por un lado la falta de acceso fácil a la información legislativa puede llevar a una menor participación ciudadana y a una comprensión limitada de los derechos y responsabilidades establecidos por la ley.

Por otro lado, los profesionales del derecho, incluidos abogados y estudiantes de derecho, dependen en gran medida de la

capacidad de acceder rápidamente a la información legislativa relevante para llevar a cabo su trabajo de manera efectiva.

Por lo tanto, mejorar la accesibilidad y la usabilidad de la información legislativa en el sitio web de la Honorable Cámara de Diputados no solo beneficiaría a los ciudadanos comunes al fortalecer la participación democrática, sino que también facilitaría el trabajo de los profesionales del derecho, promoviendo así un sistema legal más eficiente y justo para todos los argentinos.

III. LARGE LANGUAGE MODEL

En los últimos años, los modelos de lenguaje han demostrado ser herramientas poderosas en una amplia gama de aplicaciones, desde la generación de texto hasta la traducción automática y la comprensión del lenguaje natural. En el contexto de la recuperación y procesamiento de información legislativa, los LLM (Large Language Models) están emergiendo como una herramienta prometedora para mejorar la accesibilidad y la usabilidad de los datos legislativos.

Un enfoque popular es el uso de modelos de pregunta-respuesta (QA) basados en LLM, que pueden proporcionar respuestas precisas a preguntas específicas sobre la legislación. Estos modelos, están diseñados para comprender el contexto legislativo y generar respuestas basadas en la información disponible en los textos legales.

III-A. Transformers

El modelo de Transformers[5] ha revolucionado el campo de la inteligencia artificial. Su introducción marcó un hito en el desarrollo de modelos de lenguaje. A diferencia de las arquitecturas previamente usadas, los Transformers se basan en mecanismos de atención para procesar secuencias de entrada.

III-B. Retrieval Augmented Generation

Es una técnica que ha surgido para mejorar la capacidad de los modelos de lenguaje en la recuperación de información y la generación de respuestas. RAG combina dos enfoques clave: recuperación y generación.

En el primer paso, el modelo realiza una recuperación de documentos relevantes utilizando una estrategia de recuperación de información. Esto puede involucrar técnicas como la búsqueda por relevancia o la recuperación de documentos basada en vectores.

Una vez que se han recuperado los documentos relevantes, el modelo de lenguaje generará una respuesta basada en la información contenida en estos documentos.

IV. MARCO TEÓRICO

A continuación se presentan los conceptos teóricos que sustentan el proyecto

IV-A. Accesibilidad de la Información

En el contexto digital, la usabilidad se centra en la experiencia del usuario (UX), que abarca todos los aspectos de la interacción entre un usuario y un sistema o plataforma. La accesibilidad de la información es un principio fundamental en la promoción de la transparencia y la participación ciudadana en una sociedad democrática. En su esencia, implica garantizar que la información relevante esté disponible y sea comprensible para todos los ciudadanos, sin importar sus capacidades o circunstancias individuales. Este principio se basa en la idea de que el acceso a la información es un derecho humano fundamental, esencial para el ejercicio pleno de la ciudadanía y la toma de decisiones informada.

IV-B. Usabilidad en Sistemas de Información

La experiencia del usuario se ve influenciada por diversos factores, como la facilidad de navegación, la claridad de la presentación de la información, la consistencia del diseño, la capacidad de respuesta del sitio web o aplicación, y la capacidad de encontrar rápidamente la información deseada. Una interfaz intuitiva y bien diseñada puede mejorar significativamente la experiencia del usuario y fomentar la satisfacción, la lealtad y el compromiso a largo plazo.

IV-C. Modelos de Lenguaje

Los modelos de lenguaje son herramientas fundamentales en el campo del procesamiento del lenguaje natural (PLN) que permiten a las computadoras comprender y generar texto de manera inteligente. Una de las características más importantes de los modelos de lenguaje es su capacidad para comprender la semántica del lenguaje humano. Esto significa que pueden captar el significado y el contexto de las palabras y oraciones, lo que les permite interpretar el texto de manera significativa y producir respuestas relevantes y coherentes.

IV-D. Fine Tuning vs RAG

El Fine-Tuning implica ajustar un modelo de lenguaje pre-entrenado para adaptarlo a una tarea específica, utilizando datos de entrenamiento. Es efectivo para tareas específicas con datos de entrenamiento abundantes, pero puede requerir muchos recursos y ser costoso.

Por otro lado, el Método de Atención Escalonada (RAG) combina la recuperación de información y la generación de respuestas en un solo marco. Utiliza un primer paso de recuperación de documentos relevantes seguido de un segundo paso de generación de respuestas basado en el contenido de estos documentos. RAG permite una búsqueda más selectiva en grandes conjuntos de datos, mejorando la precisión y relevancia de las respuestas.

Ambos enfoques tienen sus ventajas y desventajas, para casos donde la información legislativa esta en constante modificación se prefirió utilizar el enfoque de RAG.

IV-E. Text-to-SQL

Text-to-SQL es una tarea en el campo del procesamiento del lenguaje natural que consiste en convertir consultas de lenguaje natural en consultas estructuradas de SQL que pueden ser ejecutadas en bases de datos. En los últimos años, los modelos de lenguaje de gran escala, como BERT y GPT, han demostrado ser eficaces en la realización de esta tarea.

V. COMPONENTES

En esta sección, se describirán los componentes del sistema propuesto para llevar a cabo el proyecto.

V-A. Modelo de Lenguaje

El proyecto contemplará el uso de RAG para alimentar al modelo de lenguaje de preferencia, optar por un modelo de lenguaje genérico, optimizado y potente será suficiente para responder preguntas básicas acerca de la información legislativa.

V-B. Integración de Fuentes de Datos

Mediante el uso de RAG se creará un embedding con la información que se desee alimentar al modelo de lenguaje. Una tecnología candidata en este caso es por ejemplo Chroma, que crece rápidamente con el auge en el uso de inteligencia artificial.

V-C. Backend

Utilizando un backend escrito en Python, debido a los beneficios que nos ofrece este lenguaje en el uso de inteligencia artificial, se tomará la consulta del usuario para posteriormente obtener la información del embedding previamente creado. Con esta información que obtuvimos alimentaremos al modelo de lenguaje entregándole un contexto rico en información que lo ayude a responder de manera eficiente.

V-D. Almacenamiento de Datos

Se deberá mantener almacenado el contexto de cada conversación en caso de querer que el modelo de lenguaje mantenga una conversación a través de una serie de mensajes. Un gran candidato para almacenar estos datos de manera veloz es MongoDB.

V-E. Interfaz de Usuario

La interfaz de usuario deberá ser intuitiva y veloz. Una gran tecnología candidata para facilitarnos el desarrollo de esa famosa impresión de escritura en tiempo real del modelo de lenguaje es React JS.

VI. RESULTADO

El modelo de lenguaje y las tecnologías utilizadas han demostrado ser efectivos en la comprensión del lenguaje legislativo, mejorando la precisión y relevancia de las respuestas proporcionadas a los usuarios. La interfaz de usuario diseñada promueve una experiencia más satisfactoria y efectiva, facilitando el acceso y la navegación a través de la información legislativa y fácil acceso. En resumen, estos resultados destacan el potencial de las tecnologías de procesamiento de lenguaje natural y el diseño centrado en el usuario para mejorar el acceso a la información legislativa, promoviendo la participación ciudadana.

VII. CONCLUSIÓN Y TRABAJOS FUTUROS

En esta investigación, hemos abordado la problemática del acceso a la información legislativa dentro de la Honorable Cámara de Diputados de la Nación, explorando el potencial de las tecnologías de procesamiento de lenguaje natural para mejorar la usabilidad y la accesibilidad de dicha información.

Durante el desarrollo de esta investigación, hemos enfrentado varios obstáculos y desafíos, entre ellos, el método de recolección de información. Se eligió con mucha atención el método de RAG por sobre el Fine-Tuning, debido al carácter cambiante de los datos, su posibilidad de actualización veloz y su optimización en costos y tiempo. Entre otros de los grandes desafíos se encontró la restricción de inyección y manipulación de la generación de texto. Elegir políticas de restricción y elegir con atención las instrucciones de sistema que se le debían dar al modelo de lenguaje.

Los obstáculos encontrados durante esta investigación nos brindan valiosas lecciones sobre cómo avanzar en futuros trabajos en este campo. Se recomienda para futuros trabajos, mejorar la capacidad de limitar las respuestas del modelo de lenguaje, para que este solo se comporte como un modelo que otorga información, en ningún momento deberá emitir una opinión política que pueda afectar la capacidad de decisión democrática de una persona.

En conclusión, esta investigación ha demostrado el gran potencial de las tecnologías de procesamiento de lenguaje natural para mejorar el acceso a la información de carácter público, para de esta manera poder abrirle las puertas de las políticas de cada país al mundo. Nuestros resultados sugieren que este enfoque tiene el potencial de contribuir significativamente a la política.

REFERENCIAS

- [1] "Sitio web de la hcdn," <https://www.hcdn.gob.ar/>, consultado: 2024.
- [2] "Buscador de proyectos de la hcdn," <https://www.hcdn.gob.ar/proyectos/>, consultado: 2024.
- [3] "Buscador de proyectos de la hcdn, folio," https://www.hcdn.gob.ar/folio-cgi-bin/om_isapi.dll, consultado: 2024.
- [4] A. B. Daniil, M. Robert, and G. Gabe, "Emergent autonomous scientific research capabilities of large language models," *LLM Capabilities*, 2023.
- [5] V. Ashish, J. Llion, S. Noam, P. Niki, N. G. Aidan, U. Jakob, K. Łukasz, and P. Illia, "Attention is all you need," *AI Research*, 2014.
- [6] L. Zhenwen and X. Tao, "Using llm to select the right sql query from candidates," *Text-to-SQL*, 2024.

- [7] L. Jinyang, H. Binyuan, Q. Ge, Y. Jiayi, L. Binhua, L. Bowen, W. Bailin, Q. Bowen, G. Ruiying, H. Nan, Z. Xuanhe, M. Chenhao, L. Guoliang, K. C. Chang, H. Fei, C. Reynold, and L. Yongbin, "Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls," *Text-to-SQL*, 2024.
- [8] Z. Chao, M. Yuren, F. Yijiang, M. Yu, G. Yunjun, C. Lu, L. Dongfang, and L. Jinshu, "Finsql: Model-agnostic llms-based text-to-sql framework for financial analysis," *Text-to-SQL*, 2024.
- [9] B. Yoshua, D. Réjean, V. Pascal, and J. Christian, "A neural probabilistic language model," *Probabilistic Language Model*, 2003.
- [10] T. Toni, "Vector database management systems: Fundamental concepts, use-cases, and current challenges," *Vector Databases*, 2023.
- [11] H. Cheng-Yu, L. Chun-Liang, Y. Chih-Kuan, N. Hootan, F. Yasuhisa, R. Alexander, K. Ranjay, L. Chen-Yu, and P. Tomas, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," *LLM*, 2023.
- [12] H. S, E. K, and F. H, "Fine tuning vs. retrieval augmented generation for less popular knowledge," *RAG, Fine-tuning*, 2024.
- [13] B. Angels, B. Vinamra, C. Renato, E. Roberto, H. Todd, H. Daniel, M. Jennifer, M. Nick, M. Sara, O. N. Leonardo, P. Rafael, S. Morris, S. Bruno, S. Swati, A. Vijay, and C. Ranveer, "Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture," *RAG vs Fine-tuning*, 2024.
- [14] OpenAI, "Gpt-4 technical report," *GPT-4*, 2024.
- [15] T. Ayush and G. Raghav, "Introducing super rags in mistral 8x7b-v1," *Super RAG*, 2024.