JurisGPT: Un Modelo de Lenguaje para la Generación de Sumarios de Sentencias de la Suprema Corte de Justicia de Mendoza

Rodrigo Gonzalez [a,c] y Carlos A. Catania [b,c]

[a] Universidad Tecnológica Nacional, Mendoza, Argentina

rodralez@frm.utn.edu.ar

[b] Universidad Nacional de Cuyo, Facultad de Ingeniería, LABSIN, Mendoza, Argentina

harpo@ingenieria.uncuyo.edu.ar

[c] Universidad Champagnat, Facultad de Informática y Diseño, Mendoza, Argentina

Resumen

Este proyecto de investigación, conocido como JurisGPT, se centra en la creación de un modelo de lenguaje avanzado diseñado específicamente para la generación de sumarios de sentencias emitidas por la Suprema Corte de Justicia de la provincia de Mendoza. A diferencia de los resúmenes tradicionales, el propósito fundamental de JurisGPT es extraer la jurisprudencia esencial contenida en un fallo judicial. Su principal objetivo es brindar asistencia a los relatores de la Corte Suprema en su labor diaria. Este modelo de lenguaje tiene la tarea de leer y comprender minuciosamente los extensos fallos judiciales de la Suprema Corte de Justicia, para luego sintetizar de manera concisa y precisa las partes más relevantes en un sumario. La automatización de este proceso tiene el potencial de ahorrar tiempo y esfuerzo a los relatores, eliminando la necesidad de redactar manualmente los resúmenes de las sentencias, al tiempo que se espera mejorar la eficiencia en el manejo de información legal. El acceso rápido a información esencial es una característica clave de este modelo de lenguaje, y se anticipa que mejorará tanto la eficiencia como la calidad del trabajo llevado a cabo en la Suprema Corte de Justicia de Mendoza. JurisGPT es una iniciativa conjunta entre la Universidad Tecnológica Nacional (UTN), la Universidad Nacional de Cuyo (UNCuyo), la Universidad Champagnat (UCH) y la propia Suprema Corte de Justicia de la provincia de Mendoza. Para garantizar la calidad del modelo, se llevaron a cabo pruebas exhaustivas y análisis comparativos entre diferentes sistemas de modelos de lenguaje disponibles. En el proceso de desarrollo, se evaluaron tres modelos con capacidades de 7, 13 y 70 mil millones de parámetros, utilizando la métrica ROUGE como criterio de evaluación. Los resultados preliminares revelan el potencial de los modelos de lenguaje en el ámbito jurídico en general, y su aplicación específica en este proyecto.

Palabras Clave: Modelos grandes de lenguaje, NLP, LLM, Sumarios judiciales, Inteligencia artificial

1 Introducción

Esta investigación se enfoca en el proyecto JurisGPT, una iniciativa cuyo objetivo principal radica en el desarrollo de un modelo de lenguaje avanzado diseñado para generar sumarios de sentencias emitidas por la Suprema Corte de Justicia de la provincia de Mendoza. En contraposición a los resúmenes tradicionales, el núcleo esencial de JurisGPT reside en la extracción de jurisprudencia relevante de fallos judiciales, con la finalidad de proporcionar una herramienta de apoyo fundamental para los relatores de la Corte Suprema en su labor cotidiana.

LLaMA (Large Language Model Meta AI) [1] es un modelo de lenguaje de gran envergadura lanzado por Meta AI en febrero de 2023. LLaMA ha sido entrenado en diversas escalas, que varían desde 7

mil millones hasta 65 mil millones de parámetros. Según los informes de los desarrolladores de LLaMA, el modelo de 13,000 millones de parámetros supera en rendimiento a modelos más grandes, como el GPT-3, que cuenta con 175,000 millones de parámetros. A diferencia de la tendencia previa de restringir el acceso a los modelos de lenguaje más potentes, Meta AI lanzó LLaMA a la comunidad de investigación bajo una licencia no comercial. En 18 de julio de 2023, Meta, en colaboración con Microsoft, anunció Llama 2 [2], la siguiente generación de LLaMA. Llama 2 se entrenó y lanzó en tres tamaños de modelo: 7 mil millones, 13 mil millones y 70 mil millones de parámetros. Aunque la arquitectura del modelo se mantuvo en gran medida sin cambios con respecto a los modelos LLaMA, se utilizó un 40% más de datos para entrenar los modelos fundamentales. Llama 2 incluye tanto modelos fundamentales como modelos afinados para el diálogo, denominados Llama 2 Chat. Además, en un cambio significativo en comparación con LLaMA, todos los modelos se distribuyen con sus pesos y están disponibles de forma gratuita para muchos casos de uso comercial, aunque algunas restricciones aún persisten.

En el marco de esta investigación, se utilizaron las tres versiones de Llama 2, con 7, 13 y 70 mil millones de parámetros para la generación de sumarios de las sentencias de la Corte Suprema.

La elección de SBERT (Sentence-BERT) [3] como método de embeddings desempeña un papel crucial en esta investigación. SBERT se destaca por su capacidad para capturar relaciones semánticas entre oraciones y fragmentos de texto, lo que lo hace ideal para tareas de procesamiento de lenguaje natural como la generación de resúmenes y la búsqueda de información legal. SBERT, mediante el aprendizaje siamese networks y el entrenamiento en grandes conjuntos de datos, logra representar de manera efectiva la similitud semántica entre palabras, oraciones y documentos. En el contexto de JurisGPT, esto permite la identificación precisa de jurisprudencia relevante dentro de los fallos judiciales, lo que es esencial para la generación de resúmenes precisos y contextualmente ricos. Además, SBERT es especialmente útil cuando se trata de manejar textos legales altamente estructurados y técnicos, ya que captura la semántica y la coherencia de manera más efectiva que enfoques tradicionales.

En términos de evaluación de los sumarios generados a partir de las sentencias judiciales, se recurrió a la métrica ROUGE [4][5], ampliamente reconocida en el campo de procesamiento del lenguaje natural, con el fin de medir la calidad y la coherencia de los resúmenes en comparación con los textos de referencia. Esta métrica proporciona una evaluación objetiva y cuantitativa de la eficacia del modelo JurisGPT en la extracción de jurisprudencia relevante de las sentencias de la Corte Suprema de Mendoza. La métrica ROUGE (Recall-Oriented Understudy for Gisting Evaluation) funciona comparando la similitud entre el texto generado por un modelo (por ejemplo, un sumario automático) y un texto de referencia humano (la versión correcta o de experto). ROUGE calcula la coincidencia de n-gramas (conjuntos de palabras contiguas) y evalúa la superposición entre ellos. Mide la precisión y el alcance de las palabras clave compartidas entre el texto generado y el de referencia, generando una puntuación que va de 0 a 1, la cual indica qué tan bien se aproxima el texto generado al texto de referencia, lo que proporciona una evaluación cuantitativa de la calidad del resumen generado.

Para llevar a cabo los experimentos, se emplearon varias tecnologías clave. Se utilizó Amazon SageMaker para crear notebooks de trabajo y acceder a los modelos a través de los endpoints proporcionados por AWS (Amazon Web Services).

Cabe destacar que todo el código desarrollado para este trabajo de investigación puede ser encontrado en el siguiente repositorios de archivos [6].

En las secciones siguientes, se presentan en detalle tanto el proceso de desarrollo de los experimentos como los resultados alcanzados hasta el momento. En la sección de Metodología, se describen minuciosamente los pasos que se llevaron a cabo durante la elaboración de los experimentos. Por su parte, la sección de Resultados proporciona un análisis de los resultados obtenidos. Finalmente, en la sección de Conclusiones, se resume este trabajo y se delinean las posibles direcciones para futuras investigaciones.

2 Metodología

En el ámbito de los Modelos de Lenguaje de Gran Tamaño (LLM), una de las aplicaciones más destacadas es la sumarización de texto. Esta investigación se enfoca en la aplicación de técnicas de sumarización mediante LLM, específicamente en la generación de sumarios de sentencias judiciales emitidas por la Suprema Corte de Justicia de la provincia de Mendoza.

2.1 Sumarios vs. Resúmenes

Es importante aclarar que el objetivo principal de este trabajo no consiste en crear resúmenes convencionales de fallos judiciales, sino en la generación de lo que en el ámbito legal se conoce como "sumario". Un sumario de un fallo judicial es una descripción concisa de las doctrinas legales contenidas en la sentencia, ya que una sentencia puede abordar múltiples cuestiones jurídicas. Por lo general, los sumarios son redactados por expertos de la Suprema Corte.

2.2 Pasos de la Metodología

Para obtener los sumarios de los fallos judiciales, se siguieron los siguientes pasos:

a. Creación del prompt: La creación de un prompt específico para guiar al LLM en la generación del sumario resulta fundamental. Se realizaron varias pruebas y ajustes en la formulación del prompt para obtener resultados aceptables.
b. Solicitud de resúmenes por sección: Se solicitó al LLM que generara un resumen para cada sección del fallo, utilizando un prompt específico diseñado para esta tarea. Dado que Llama 2 tiene una ventana de contexto limitada a 4096 tokens, se dividió el fallo en las secciones que lo componen, asegurando así que el modelo pudiera procesar cada parte de manera efectiva. Además, si la sección superaba los 4096 tokens de extensión fue necesario dividirla en partes (chunks) y solicitar al LLM que hiciera un resumen de los pedazos de textos proporcionados.
c. Solicitud de creación de un sumario: finalmente, con los resúmenes por cada sección, se le pide al LLM que proporcione un sumario del fallo. Se debe crear un nuevo prompt para este propósito.
d. Evaluación del sumario mediante ROUGE: El sumario generado por el LLM se compara con el sumario elaborado por un experto de la Suprema Corte de Mendoza utilizando la métrica ROUGE. Esta métrica se utiliza para medir la similitud y calidad de los textos generados en comparación con el texto de referencia, proporcionando una evaluación objetiva de la eficacia del modelo en la extracción de jurisprudencia relevante.

3 Resultados

En total, se analizaron 10 fallos de la Suprema Corte en el ámbito Laboral utilizando los tres modelos de Llama 2 previamente mencionados: Llama 2 7B, 13B y 70B.

Para ilustrar el rendimiento de estos modelos, la Tabla 1 presenta una comparación entre el sumario generado por Llama 2 70B para un fallo específico y el sumario elaborado por un experto de la Suprema Corte.

Tabla 1. Comparación del sumario provisto por Llama 2 70B respecto el sumario escrito por un experto de la Suprema Corte, para una fallo en particular.

Sumario del LLM	Sumario de un experto
En resumen, la Sentencia de la Corte Suprema de Mendoza, declara la nulidad del despido del trabajador de la cooperativa, debido a la falta de participación de los asociados en la dirección y vida de la cooperativa, y a la ausencia de presentación de balances para su aprobación en el acto asambleario pertinente. La Corte también	Cuando se trata de cooperativas no se puede pasar por alto los motivos fundados en el esfuerzo propio y la ayuda mutua para organizar y prestar servicios; su forma de constitución, condiciones de ingreso la administración, gobierno y representación legal; lo excedentes repartibles destinados a ser distribuidos en

estableció que la Cooperativa es una entidad debidamente inscripta ante los organismos de control, lleva la contabilidad en forma legal y cumple con las normas de la cooperativa.	concepto de retorno entre los asociados en proporción al trabajo efectivamente prestado por cada uno de éstos. Y que los asociados a las cooperativas de trabajo no revisten la calidad de dependientes de las mismas, debiéndoselos considerar como trabajadores autónomos.

En la Tabla 1, es evidente que el modelo Llama 2 ha generado efectivamente un resumen del fallo en lugar de un sumario, lo que sugiere diferencias significativas en el enfoque de ambas aproximaciones. Esta situación se repite con el resto de los sumarios generados por los 3 modelos utilizados.

A continuación, en la Tabla 2, se resumen los promedios y las desviaciones estándar de las métricas ROUGE para los tres modelos de lenguaje al procesar los 10 fallos de la corte en cada caso.

Tabla 2. Valores estadísticos de las métricas ROUGE para los 3 modelos bajo estudio al procesar 10 fallos

Modelo del lenguaje	Media	Desviación estándar
Llama 2 7B	0.2739	0.0546
Llama 2 13B	0.2885	0.0656
Llama 2 70B	0.2913	0.0739

La Tabla 2 revela una ligera mejora a medida que se incrementan los parámetros del modelo Llama 2. Sin embargo, se considera que esta mejora es relativamente insignificante en el contexto de los resultados obtenidos.

4 Conclusiones

Este estudio se ha centrado en el proyecto JurisGPT, una iniciativa que tiene como propósito principal el desarrollo de un modelo de lenguaje avanzado específicamente diseñado para generar sumarios de sentencias emitidas por la Suprema Corte de Justicia de la provincia de Mendoza. Se utilizaron tres modelos de la familia Llama 2, con 7B, 13B y 70B parámetros, para evaluar su capacidad en la generación de estos sumarios.

A pesar de los esfuerzos realizados en la iteración del prompt para la generación de sumarios finales de fallos judiciales, los resultados en términos de las métricas ROUGE, que se acercan a 0.3, indican que aún queda margen para mejorar el sistema de generación de sumarios automáticos. Estos valores sugieren que el modelo, en su estado actual, no ha logrado generar sumarios que se equiparen a la calidad de los elaborados por expertos humanos.

Un hallazgo importante es que el aumento de los parámetros del modelo Llama 2 no proporciona una mejora notable en la calidad de los sumarios generados. Esto sugiere la necesidad de explorar otras estrategias de mejora, como ajustes específicos en el entrenamiento o en la estructura del modelo.

Como una dirección prometedora para futuras investigaciones, se propone llevar a cabo un ajuste fino (fine-tuning) del modelo Llama 2 7B. Este ajuste se basará en datos de entrada que incluirían tanto los sumarios generados por un modelo Llama 2 70B como los sumarios creados por expertos de la Suprema Corte. Este enfoque podría proporcionar una oportunidad para afinar y calibrar el modelo de manera más precisa, aspirando a una mejora significativa en la calidad de los sumarios generados automáticamente.

En resumen, este estudio inicialmente valioso destaca la necesidad continua de perfeccionar los modelos de lenguaje avanzado para aplicaciones jurídicas específicas como la generación de sumarios de sentencias judiciales. A través del aprendizaje de los resultados y la implementación de futuras investigaciones, se espera que JurisGPT y proyectos similares puedan contribuir de manera más efectiva al trabajo en el ámbito legal y la mejora de la eficiencia en la Suprema Corte de Justicia de Mendoza.

Agradecimientos

Los autores de este trabajo desean expresar su sincero agradecimiento por el valioso apoyo brindado por los miembros del equipo de la Suprema Corte de la provincia de Mendoza.

Referencias

[1] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[2] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).

[3] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[4] Lin, Chin-Yew, and F. J. Och. "Looking for a few good metrics: ROUGE and its evaluation." Ntcir workshop. 2004.

[5] Barbella, Marcello, and Genoveffa Tortora. "ROUGE metric evaluation for Text Summarization techniques." Available at SSRN 4120317 (2022).

[6] Rodrigo Gonzalez. JurisGPT: an AI-powered Summarization System for the Supreme Court Rulings of the Mendoza province, Argentina. URL: https://github.com/rodralez/jurisgpt