Título de la Tesis:"Desarrollo de un Sistema para la Interacción Efectiva entre Modelos de Lenguaje y Datos Estructurados en Anuarios Estadísticos"

Carla Sunami Pérez Valera

24 de diciembre de 2024

Introducción

El procesamiento de datos ha evolucionado significativamente en las últimas décadas, impulsado por el crecimiento exponencial de la información disponible y la necesidad de análisis precisos en un mundo cada vez más complejo, donde la capacidad de extraer, analizar y utilizar datos de manera efectiva se ha convertido en un pilar para la toma de decisiones en múltiples disciplinas. En este contexto, los anuarios estadísticos juegan un papel crucial al proporcionar datos sistemáticos sobre diversos aspectos económicos, sociales y demográficos, siendo un compendio valioso que reflejan la evolución y el estado actual. En Cuba, la Oficina Nacional de Estadística e Información (ONEI) es responsable de compilar y publicar estos anuarios.

Históricamente, Cuba ha enfrentado desafíos económicos significativos que han influido en su desarrollo social y político. La recopilación y análisis de datos a través de anuarios estadísticos permiten a los investigadores y responsables políticos comprender mejor las dinámicas económicas del país y tomar decisiones basadas en evidencia. Sin embargo, el acceso a esta información no siempre se traduce en una comprensión inmediata. La complejidad y el volumen de datos presentados en estos anuarios pueden dificultar su interpretación.

Si bien los datos tabulares son omnipresentes, las tareas específicas relacionadas con las tablas pueden ser laboriosas, propensas a errores y requerir habilidades especializadas. La automatización de estas tareas ofrece beneficios significativos tanto para los sectores académicos como industriales, atrayendo un interés considerable. Los métodos convencionales para procesar datos tabulares se enfocan predominantemente en adaptar arquitecturas de modelos de lenguaje, incorporando elementos como incrustaciones de posición, mecanismos de atención y objetivos de aprendizaje para codificar los atributos estructurales inherentes de los datos tabulares. Sin embargo, se ha producido un cambio de paradigma con el surgimiento de grandes modelos

de lenguaje (LLM) como GPT-4, GPT-3.5, PaLM2, Llama, entre otros. Las investigaciones recientes enfatizan la creación de indicaciones precisas que integren información parcial crucial a partir de datos tabulares proporcionados y el aprovechamiento de lenguajes de programación externos como SQL y Python.[**1**]

Los modelos de lenguaje grandes (LLM) son modelos de aprendizaje profundo entrenados con una gran cantidad de datos, lo que los dota de capacidades versátiles de resolución de problemas que se extienden mucho más allá del ámbito de las tareas de procesamiento del lenguaje natural (PLN). Investigaciones recientes han revelado capacidades emergentes de los LLM, como un mejor desempeño en tareas con pocas indicaciones. El desempeño notable de los LLM ha despertado interés tanto en el ámbito académico como en la industria, lo que ha generado la creencia de que podrían servir como base para la Inteligencia Artificial General (AGI) de esta era. Un ejemplo notable es ChatGPT, diseñado específicamente para participar en conversaciones humanas, que demuestra la capacidad de comprender y generar texto en lenguaje humano.[**2**]

Antes de los LLM, los investigadores han estado investigando formas de integrar datos tabulares con redes neuronales para tareas de PLN y gestión de datos. Hoy, los investigadores están interesados en investigar las capacidades de los LLM al trabajar con datos tabulares para diversas tareas, como predicción, comprensión de tablas, razonamiento cuantitativo y generación de datos. [**2**]

Generar características de columna adecuadas, incluso con conocimiento del dominio, puede ser desafiante y costoso. Por ejemplo, la validación manual para identificar características útiles es inviable debido a la cantidad exponencial de combinaciones posibles para explorar. Para abordar este problema, los métodos de ingeniería de características existentes utilizan esquemas de filtrado adicionales para evaluar y seleccionar características útiles de forma automática. Si bien estos enfoques reducen el esfuerzo manual y mejoran la calidad de las características, aún presentan varios desafíos. En primer lugar, los profesionales a menudo dependen de espacios de búsqueda definidos manualmente para generar características candidatas debido a la ambigüedad inherente de lo que constituye características informativas particularmente a medida que el número de características y la complejidad del espacio de búsqueda crecen. Además, descuidan los diseños experimentales más efectivos, confiando únicamente en las puntuaciones de validación para seleccionar buenas características, a pesar del valor de los datos de experimentos ante-

riores para mejorar la selección. [**3**]

Motivados por esto, proponemos abordar este problema desde una perspectiva novedosa: la optimización para descubrir reglas de generación efectivas, aprovechando la comprensión del lenguaje y las capacidades de razonamiento de los modelos de lenguaje grandes (LLM). Investigaciones recientes han demostrado que los LLM pueden optimizar varios problemas no diferenciables utilizando indicaciones que describen la tarea de optimización en lenguaje natural. Esto sugiere el potencial de los LLM para generar automáticamente y refinar iterativamente generadores de características sin la necesidad de especificar manualmente el espacio de reglas. Por ejemplo, las capacidades de razonamiento de los LLM permiten incorporar retroalimentación sobre sus resultados anteriores en el proceso de refinamiento iterativo. Además, los contextos lingüísticos, como los nombres de las columnas (por ejemplo, "Género" y "Edad") y los valores categóricos (por ejemplo, "Mujer" y "Hombre"), podrían integrarse naturalmente en la optimización, lo que es difícil, si no imposible, con los métodos convencionales.[**3**]

La mejora del rendimiento de los modelos de lenguaje grandes (LLM) en la formulación de preguntas y respuestas (QA) específicas de un dominio ha sido un foco de investigación, empleando predominantemente dos enfoques: el ajuste fino específico del dominio (DSFT), que implica el entrenamiento de los LLM en el corpus específico del dominio, y la generación aumentada por recuperación (RAG), que utiliza un corpus específico del dominio como base de conocimiento externa. Estos enfoques, que aprovechan las fortalezas inherentes del procesamiento de texto de los LLM, se han adoptado ampliamente en escenarios de solo texto, lo que ha producido mejoras significativas.[**4**]

Sin embargo, los datos del mundo real en muchos dominios suelen existir en un formato híbrido, que comprende no solo texto sino también volúmenes sustanciales de tablas semiestructuradas, como se observa, por ejemplo, en la literatura científica, los informes médicos y en nuestro caso específico en los anuarios estadísticos. Estas tablas aparecen con frecuencia junto con el texto dentro del mismo documento, lo que proporciona información semánticamente suplementaria o complementaria crucial para una comprensión integral del contenido. Al explorar el potencial de aprovechar los datos híbridos para mejorar el rendimiento de los LLM, es crucial integrar eficazmente estos datos, asegurando la coexistencia de texto y tablas. Los métodos actuales para manejar la heterogeneidad de textos y tablas tienen inconvenientes significativos:

1. Aplanar directamente las tablas mediante la concatenación de celdas

fila por fila que no solo da como resultado la pérdida de información estructural incorporada en la tabla original, sino que también corta los vínculos informativos al asignar texto y tablas a diferentes espacios vectoriales por separado y luego integrarlos, no solo aumenta la complejidad, sino que también altera la conexión semántica entre los dos tipos de datos.

2. Una solución prometedora es la generación de tablas a texto que tiene como objetivo generar declaraciones en lenguaje natural que describen fielmente la información en la tabla proporcionada. A través de esto, podemos transformar datos híbridos en una representación unificada en lenguaje natural que es más adecuada para su uso por parte de los LLM, al mismo tiempo que preserva la información importante de las tablas y las conexiones semánticas entre los datos.

Aunque la generación de tablas a texto ha sido ampliamente estudiada por la comunidad de NLP, actualmente no existe un análisis comparativo sobre cómo los corpus generados por diferentes métodos de tabla a texto afectan el desempeño de los sistemas de control de calidad específicos del dominio.[**4**]

Los documentos que contienen tanto tablas como texto, por ejemplo, presentaciones ante la SEC, artículos académicos e informes médicos, constituyen una categoría de contenido muy frecuente en el mundo real. A menudo presentan datos numéricos extensos tanto en el contenido tabular como en el textual, lo que requiere capacidades de razonamiento discreto para que las máquinas los comprendan. Investigaciones recientes investigan la comprensión inteligente de dichos documentos a través de tareas de preguntas y respuestas (QA). El modelo, provisto de una tabla y el texto relevante como contexto necesita realizar varios tipos de razonamiento discreto, como cálculos aritméticos, hacer comparaciones y contar, para responder la pregunta. Para realizar el control de calidad sobre datos textuales y tabulares híbridos, un enfoque sencillo implica tomar la tabla, el texto y la pregunta como entrada y generar la respuesta directamente. Este enfoque puede ser ineficaz debido al complejo proceso de razonamiento involucrado. Para abordar este problema, algunos trabajos descomponen la tarea en múltiples pasos, produciendo resultados intermedios que sirven como referencias para la respuesta final. Estos enfoques de múltiples pasos generalmente diseñan módulos distintos en cada paso y, a menudo, optimizan estos módulos simultáneamente a través del aprendizaje de múltiples tareas. Hasta la fecha, no ha habido consenso sobre cómo descomponer el proceso de respuesta en la literatura existente.[**5**]

Recientemente, los modelos de lenguaje grandes (LLM) como GPT-4 y FLAN han mostrado fuertes capacidades de razonamiento de múltiples pasos con instrucciones adecuadas como la cadena de pensamiento (CoT), por lo tanto, consideramos aprovechar este asombroso poder de los LLM para un mejor razonamiento discreto sobre datos híbridos tabulares y textuales. Para lograrlo, primero identificamos pasos clave en el proceso de control de calidad tabular y textual de métodos de varios pasos anteriores, y abstraemos una secuencia de pasos.[**5**]

Los pasos enfatizan diferentes capacidades del modelo de control de calidad tabular y textual: comprender la pregunta y el contexto, inferir la lógica para responder la pregunta y calcular la respuesta con precisión. Estos pasos producen una secuencia de resultados intermedios, lo que significa que podemos modelar y mejorar específicamente uno (o más) de ellos dado un escenario de aplicación específico. Sin embargo, utilizar un LLM en línea presenta desafíos en términos de costo, latencia y riesgo de seguridad de datos. [**5**]

A pesar de la disponibilidad de datos en los anuarios estadísticos, existe una brecha significativa entre la recopilación de estos datos y su utilización efectiva para responder preguntas específicas. Los Modelos de Lenguaje de Gran Escala(LLM) han demostrado ser herramientas poderosas para el procesamiento del lenguaje natural, pero su rendimiento puede verse limitado cuando se enfrentan a datos tabulares y textuales sin un contexto adecuado. Esto se debe a que los LLM generalmente se entrenan con texto no estructurado. Esto plantea una necesidad crítica: desarrollar un sistema que no solo procese estos datos, sino que también facilite la interacción con un LLM para generar respuestas precisas y relevantes.

La justificación para esta investigación radica en la necesidad de mejorar la capacidad de los LLM para responder preguntas específicas sobre datos tabulares y textuales al proporcionar un marco estructurado que facilite la interacción entre el modelo y los datos. Este enfoque no solo beneficiará a los investigadores que buscan respuestas rápidas y precisas, sino que también mejorará la accesibilidad de la información para los responsables políticos y usuarios en general.

La problemática central se relaciona con la dificultad que enfrentan los modelos de lenguaje al intentar interpretar y responder preguntas basadas en datos tabulares extraídos de anuarios estadísticos. Aunque estos modelos son capaces de procesar texto no estructurado con gran eficacia, su rendimiento disminuye cuando se trata de datos organizados en tablas. Esto limita su

utilidad en contextos donde se requiere información precisa y contextualizada sobre indicadores económicos.

Además, el volumen y la complejidad de los datos presentados en los anuarios pueden resultar abrumadores para los usuarios que no están familiarizados con el análisis estadístico. Por lo tanto, es esencial desarrollar un sistema que no solo procese estos datos, sino que también proporcione un contexto adecuado para que el LLM pueda generar respuestas relevantes.

En el contexto actual, donde los LLM están ganando popularidad en diversas aplicaciones, desde asistentes virtuales hasta análisis predictivos, la necesidad de integrar estos modelos con datos estructurados es más relevante que nunca. La novedad científica de esta investigación radica en su enfoque interdisciplinario que combina técnicas de procesamiento del lenguaje natural con análisis estadístico.

Desde una perspectiva teórica, este estudio contribuirá al campo del procesamiento del lenguaje natural al explorar cómo los LLM pueden ser mejorados mediante el uso de datos tabulares y textuales estructurados. Desde una perspectiva práctica, facilitará el acceso a información económica crítica para investigadores y responsables políticos, permitiendo una toma de decisiones más informada basada en evidencia.

La hipótesis central es que "la integración efectiva de datos tabulares y textuales extraídos de anuarios estadísticos con un modelo de lenguaje mejorará significativamente la precisión y relevancia de las respuestas generadas por dicho modelo ante consultas específicas relacionadas con indicadores económicos".

Diseño Teórico

1. Problema Científico: ¿Cómo puede un LLM ser optimizado para responder preguntas sobre datos tabulares y textuales extraídos de anuarios estadísticos?

2. Objeto de Estudio: Datos tabulares y textuales extraídos de anuarios estadísticos publicados por la ONEI.

3. Objetivos:

- Objetivo General: Desarrollar un sistema que procese tanto los datos tabulares como los textos extraídos de anuarios estadísticos para mejorar las respuestas generadas por un LLM.

- Objetivos Específicos:

- Extraer y limpiar tanto los datos tabulares como el texto desde anuarios estadísticos.

- Estructurar estos datos en una base de datos accesible.

- Enriquecer el contexto del LLM mediante metadatos.

- Evaluar la precisión del modelo al responder preguntas específicas basadas en estos datos.

4. Campo de Acción: Este estudio se centrará en el ámbito del procesamiento del lenguaje natural aplicado a análisis económicos utilizando herramientas tecnológicas contemporáneas.