Reduciendo la Huella de la IA en la Estadística Oficial: un Caso Práctico sobre Clasificación Económica

Jorge Paz-Ruza[1], Adrián Pérez-Bote[2], Roi Santos-Ríos[1], and Jesús Vilares[1]

[1] Universidade da Coruña – Centro de Investigación en TIC (CITIC), A Coruña, España
j.ruza@udc.es roi.santos.rios@udc.es jesus.vilares@udc.es
[2] Instituto Nacional de Estadística (INE) – Subdirección General de Metodología y Diseño de Muestras (SGMDM), Madrid, España
adrian.perez.bote@ine.es

Abstract

El INE y el CITIC colaboran en el proyecto CIDMEFEO para desarrollar soluciones sostenibles para la Estadística Oficial, tales como el uso de datos sintéticos para entrenamiento y el empleo de modelos locales abiertos y de menor tamaño. Presentamos como caso de uso la clasificación automática de la actividad económica de empresas y autónomos.

1 Estadística Oficial e IA

Desde hace tiempo, las organizaciones de estadística oficial estudian cómo emplear las técnicas de Inteligencia Artificial (IA) en sus procesos. Su objetivo es mejorar la comprensión de los datos y la calidad de los resultados, e incrementar la eficiencia de dichos procesos a la vez que reducir su coste. Sin embargo, su empleo implica también riesgos, sean para la privacidad y protección de los datos, la posible aparición de sesgos durante el entrenamiento y, finalmente, que se puedan requerir grandes recursos computacionales con un elevado consumo de energía, lo cual puede tener un impacto perjudicial en el medio ambiente.

Recientemente, estos organismos se han visto también afectados por la irrupción de una tecnología tan disruptiva como es la IA Generativa. Esta tecnología, basada en *transformers*, emplea Modelos de Lenguaje Grandes (LLM, por *Large Language Models*) y ofrece aún mayores capacidades, si bien acompañadas también de mayores riesgos. No sólo se acusan los riesgos ya antes comentados, sino que debemos sumarles la posible aparición de alucinaciones (*hallucinations*), que se producen cuando el modelo genera información incorrecta, falsa o sin sentido, pero la presenta de forma convincente como si fuera real. No obstante, nuestro interés se centra en su sostenibilidad económica y medioambiental, la cual ha levantado serias dudas por el altísimo coste de las infraestructuras para su ejecución y su enorme consumo eléctrico.

2 Codificación Automática en el Proyecto CIDMEFEO

Dentro del proyecto *"Ciencia e ingeniería de datos para la mejora de la función estadística oficial (CIDMEFEO)"*, el CITIC desarrolla varias líneas de investigación en el ámbito de la Estadística Oficial junto con el INE y otras universidades. El caso de la línea de *"Codificación automática con técnicas de machine learning"*, su objetivo es desarrollar sistemas de codificación automática o de soporte a la codificación para la asignación de códigos de clasificación estandarizados a respuestas en texto abierto, empleando para ello técnicas de Procesamiento del Lenguaje Natural (PLN). Actualmente nuestros esfuerzos se centran en la asignación de categorías CNAE a partir de la descripción textual que el empresario aporta, en sus propias palabras, acerca de su actividad económica.

La CNAE (Clasificación Nacional de Actividades Económicas) es una clasificación estándar de la actividad económica de las empresas. Consta de una estructura jerárquica en cuatro niveles: *sección*, el más general, representado por una letra (22 categorías); *división*: un nivel más específico, identificado por dos dígitos (87 categorías); *grupo*: tres dígitos, los dos primeros corresponden a su división (287 categorías); *clase*, nivel más detallado, cuatro dígitos donde los tres primeros corresponden a su grupo (664 categorías). Bajo un mismo epígrafe de la CNAE se agrupan distintas actividades de naturaleza similar, es decir, actividades que comparten un proceso productivo común. De este modo, actividades similares poseen códigos similares.

En el caso de nuestro proyecto, las soluciones planteadas se han regido desde un principio por la sostenibilidad económica y medioambiental:

SLM abiertos. El coste asociado a los LLM puede reducirse a márgenes razonables empleando en su lugar Modelos de Lenguaje Pequeños (SML, por *Small Language Models*), modelos de menor tamaño y potencia que los LLM pero suficientes para tareas especializadas de alcance más limitado, como es nuestro caso. Su coste económico y medioambiental es también mucho más reducido: un servidor equipado con un par de GPU de gama alta puede bastar para su despliegue. Asimismo, el empleo de modelos abiertos y la posibilidad de usar infraestructura local propia permite no depender de terceros, eliminar costes adicionales y mantener un mayor control sobre el sistema de cara a la protección y privacidad de los datos.

Aumento de datos y datos sintéticos. El tamaño del conjunto de datos reales necesarios para el *fine-tuning* del modelo puede reducirse sustancialmente empleando estas técnicas. El *aumento de datos* (*Data Augmentation*) permite aumentar la diversidad de los datos de entrenamiento creando copias ligeramente modificadas de los datos originales ya existentes, o bien generando directamente datos sintéticos completamente nuevos. El objetivo es múltiple. Primero, reducir costes económicos, temporales y energéticos (algunos de estos estudios requieren miles o decenas de miles de consultas). Segundo, mejorar el rendimiento del modelo reduciendo el desequilibrio entre clases y la escasez de datos (*data sparsity*). En nuestro caso se trata, además, de un problema inevitable por la propia naturaleza de la economía española: mientras algunos sectores económicos engloban miles de empresas, otros apenas contabilizan con un puñado. Paralelamente, el empleo de datos sintéticos, es decir irreales, aumenta la privacidad al reducir la posibilidad de revelar inadvertidamente datos sensibles reales. Asimismo, el proceso de generación es guiado por la propia guía de referencia del CNAE, en el cual se describen en detalle cada una de las categorías así como sus excepciones, permitiendo un mayor control de las alucinaciones.

Optimización de modelos. Finalmente, estamos valorando el empleo de técnicas de cuantización o precisión reducida para reducir el tamaño del modelo y, por tanto, reducir los requerimientos computacionales y el consumo asociado a su ejecución.

Agradecimientos: Trabajo parcialmente financiado por el INE (CIDMEFEO); por el MICIU/AEI y FEDER/UE (proyectos PID2023-147129OB-C21 y PID2023-147404OB-I00); por el Ministerio para la Transformación Digital y de la Función Pública y Next-GenerationEU/PRTR (TSI-100925-2023-1); y por la Xunta de Galicia (ED431C 2024/02 y ED431C 2022/44). El CITIC, acreditado como "Centro de excelencia" y "Miembro de la Red CIGUS" para el período 2024-2027, está financiado por la Xunta de Galicia y la UE a través del programa operacional FEDER Galicia 2021-27 (ED431G 2023/01).