ANÁLISIS DE LOS CLASIFICADORES DE PROGRAMACIÓN DE LENGUAJE NATURAL BASADOS EN EL MODELO LATXA-7B SOBRE TEXTOS SINTÉTICOS DE TUMORES ÓSEOS

Calvo Lorenzo Isidoro
Servicio de Cirugía Ortopédica y Traumatología. Hospital Universitario Galdakao-Usansolo (Vizcaya)

Una **LLM** (Large Language Model) es un modelo de lenguaje entrenado con grandes cantidades de texto para entender, generar y procesar lenguaje natural humano. Sirve para tareas como clasificación de textos, traducción, responder preguntas y generar conversaciones. En este trabajo preentamos la potencial utilidad del LLM Latxa-7b a la hora de clasificar textos médicos en euskera.

Material y métodos

Se crea una database de 20.000 notas clínicas sintéticas de pacientes con patologías musculoesqueléticas, de las cuales se seleccionan 5.000 para entrenamiento y testeo. Para la tokenización de las historias clínicas se utilizará el LLM Latxa-7b. Una vez comprobado su óptimo funcionamiento, se procede a tokenizar e indexar las 5000 notas clínicas que se utilizarán para generar el clasificador. Las 5000 notas clínicas seleccionadas se dividirán aleatoriamente en dos grupos, uno de entrenamiento (90%) y otro de test (10%). El grupo de entrenamiento se utilizará para entrenar una red neuronal convolucional (DCNN). El modelo clasificador que resulte será testado con las notas clínicas del grupo de test.

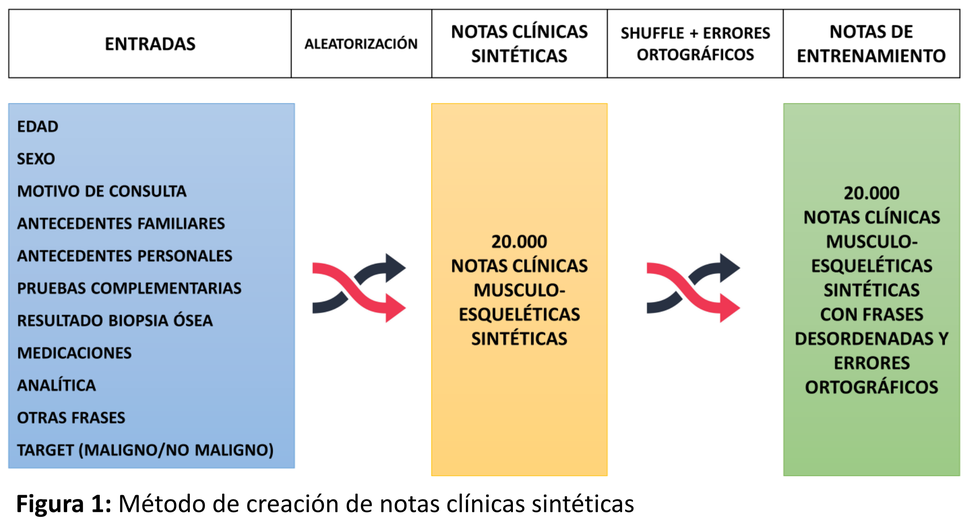

Figura 1: Método de creación de notas clínicas sintéticas

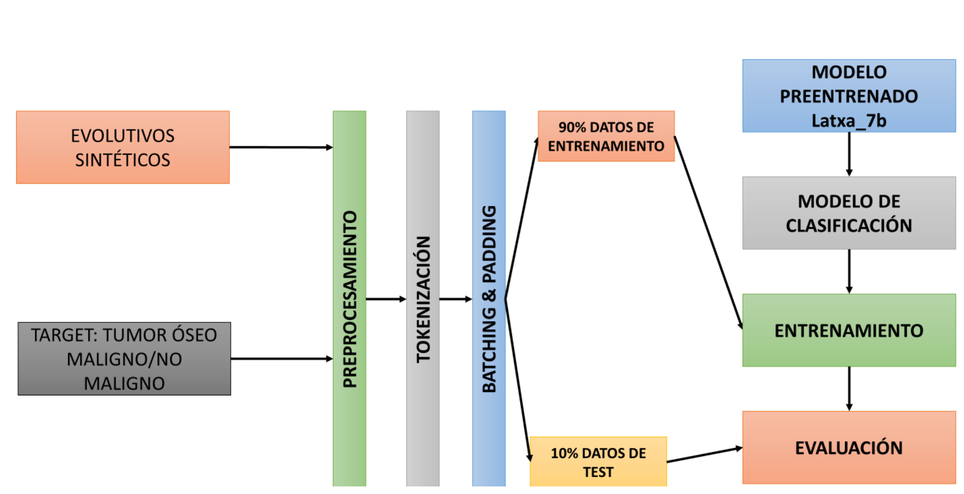

Figura 2: Estrategia de diseño del clasificador

Resultados

Se genera de este modo un modelo de clasificador cuyo rendimiento en el grupo de entrenamiento es de una exactitud del 97,7%, precisión del 98,6%, recall del 94,2%, Area bajo la Curva de 0,99 y un F1 de 0,96. Cuando se aplica al clasificador el grupo de test, los resultados son muy similares: exactitud del 97,7%, precisión del 98,6%, recall del 94,2%, Area bajo la Curva de 0,99 y un F1 de 0,96. Estos resultados indican que no hay sobreajuste, por lo cual se confirma que el modelo ha aprendido y no memorizado de los datos suministrados.

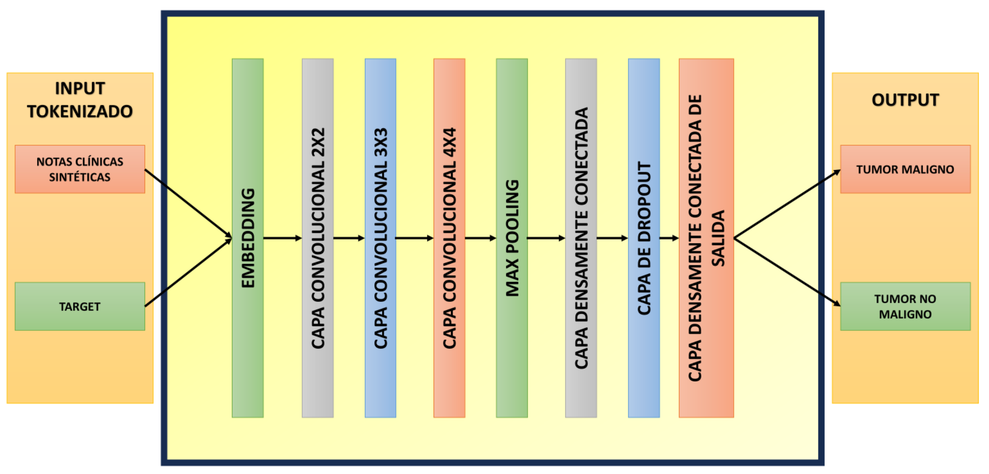

Figura 3: Modelo de entrenamiento

Conclusiones

A pesar de haber trabajado con datos sintéticos y de que se trata de un modelo en fase inicial de desarrollo, todo parece apuntar que el LLM Latxa va a permitir aplicar en el futuro estas tecnologías en textos en euskera. Su excelente rendimiento como base para un clasificador de textos médicos debería ser un acicate para implementar las LLMs y el procesamiento de lenguaje natural en las historias clínicas digitalizadas que utilizamos en nuestros sistemas sanitarios.

61 CONGRESO SECOT
SOCIEDAD ESPAÑOLA DE CIRUGÍA ORTOPÉDICA Y TRAUMATOLOGÍA

REFERENCIA: Calvo-Lorenzo I. Latxa-7b ereduan oinarritutako hizkuntzaren prozesamendu sailkatzaileen gaitasunaren azterketa: medikuntzako aplikazioak eta kirurgia ortopediko eta traumatologiako testu klinikoen adibidea. Gac Med Bilbao. 2024;121(2):62-68