Aplicación de BioBERT y RAG para la Clasificación y Consulta de Literatura Científica en el Contexto de la Pandemia de COVID-19: Un Estudio Basado en el CORD-19

Silvia Maria Gamarra Morel

Resumen

Este artículo presenta un enfoque para el análisis de la literatura científica relacionada con el COVID-19, utilizando el corpus *Open Research Dataset Challenge* (CORD-19). Inicialmente, se contempló entrenar dos modelos *BERT* preentrenados en literatura científica y biomédica, SciBERT y BioBERT. Sin embargo, debido a limitaciones en la capacidad computacional, se optó por utilizar únicamente el modelo BioBERT, que está específicamente optimizado para procesar datos biomédicos, el cual se ajusta más estrechamente con la naturaleza del conjunto de datos seleccionado.

Para llevar a cabo el estudio, se realizó *fine-tuning* de tres variantes de BioBERT para la clasificación binaria de artículos científicos, seleccionando el modelo con el mejor desempeño según las métricas de evaluación. Una vez identificado el modelo más eficiente, se almacenó una copia en un repositorio para su uso posterior. A continuación, se implementó un sistema de **Recuperación de Respuestas Generativas** (*Retrieval-Augmented Generation*, RAG), que optimiza la búsqueda y consulta de textos dentro del corpus.

El propósito de esta investigación es desarrollar una herramienta efectiva que permita extraer información relevante sobre el COVID-19 de un corpus en continuo crecimiento. Los resultados obtenidos demuestran que el enfoque propuesto no solo mejora la precisión en la clasificación, sino también la calidad de las respuestas generadas durante las consultas.

Planteamiento del problema

Durante la pandemia de COVID-19, se generó literatura científica sin precedentes, en un intento global por comprender el virus, sus mecanismos de transmisión, tratamientos posibles y efectos a largo plazo. El conjunto de datos CORD-19 reúne esta vasta producción científica, convirtiéndose en un recurso valioso para analizar cómo respondió la comunidad científica y médica ante una crisis sanitaria global.

Sin embargo, debido al volumen, heterogeneidad y naturaleza técnica de los textos, extraer lecciones útiles de esta información sigue siendo un reto; por lo cual existe la necesidad de desarrollar modelos de comprensión de lenguaje natural que permitan analizar retrospectivamente esta literatura para responder preguntas como: ¿Cómo evolucionaron las estrategias de tratamiento a lo largo del tiempo? ¿Qué factores se asociaron con la aceptación o el rechazo de ciertos enfoques terapéuticos?

Tradicionalmente, las búsquedas se realizan mediante palabras clave o métodos de clasificación sencillos, pero estas técnicas no siempre permiten una extracción precisa y relevante de los textos.

Este trabajo aborda la necesidad de desarrollar un sistema avanzado que combine técnicas de procesamiento de lenguaje natural (PLN) para clasificar los artículos y permitir consultas de manera eficiente.

Descripción del corpus

El corpus utilizado en este estudio es el *COVID-19 Open Research Dataset Challenge (CORD-19),* un conjunto de datos en constante expansión que incluye artículos científicos, preprints, y artículos relacionados con el COVID-19. Este conjunto de datos es mantenido por el *Allen Institute for AI*, en colaboración con diversas instituciones científicas. Contiene más de 200.000 artículos que cubren una amplia variedad de temas relacionados con la pandemia, el SARS-CoV-2, y otros coronavirus. Los artículos están escritos en su mayoría en inglés, aunque algunos pueden contener idiomas adicionales. En este trabajo, se utiliza un subconjunto de este corpus para entrenar y evaluar los modelos.

Metodología

Modelos utilizados

Se emplean dos modelos clave en este estudio:

1. **BioBERT v1.1**: es una versión preentrenada de BERT, optimizada para tareas biomédicas y científicas. Se entrenó desde BERT-base, pero exclusivamente con PubMed, durante 1 millón de pasos, lo que lo hace más especializado que BioBERT v1.0, su predecesor.

   BioBERT v1.1 fue desarrollado por el equipo de investigación del DMIS Lab (*Deep Learning for Medical Information Search*) del Korea University, en colaboración con NAVER Corporation, una de las principales compañías tecnológicas de Corea del Sur.

Según referencias bibliográficas es el más probado y estable a la fecha. Hay una versión nueva, la v1.2, lanzada en octubre en el 2024, más flexible y con capacidad de generar texto biomédico, pero como en este proyecto solo se quiere identificar temas, tipos de estudios, hallazgos médicos e investigaciones hechas referentes al COVID-19, con la versión 1.1 bastará.

Para este trabajo, se fine-tunea BioBERT en un conjunto de datos etiquetado para realizar una clasificación binaria de los artículos del CORD-19, artículos COVID vs. no COVID.

2. **RAG (*Retrieval-Augmented Generation*)**: el modelo RAG se utiliza para mejorar la capacidad de consulta del sistema. RAG combina modelos de recuperación de documentos con modelos generativos (como GPT), lo que permite generar respuestas a consultas basadas en los artículos recuperados.
   El modelo RAG utiliza un componente de recuperación para identificar los documentos más relevantes dentro del corpus dado (en este caso, artículos del CORD-19).
   En el contexto de este trabajo, RAG se implementa para optimizar la búsqueda de artículos dentro del corpus CORD-19, permitiendo generar respuestas detalladas a consultas sobre temas como tipos de estudios, hallazgos médicos y avances en investigaciones relacionadas con el COVID-19. La combinación de recuperación y generación mejora la precisión y relevancia de las respuestas generadas, proporcionando a los investigadores una herramienta más efectiva para extraer información valiosa de la vasta cantidad de literatura disponible.

Procesamiento de datos

El procesamiento de los datos para el entrenamiento del modelo BioBERT v1.1 con *fine-tuning, se* incluyó diversas etapas fundamentales para preparar el corpus CORD-19 con fines de entrenamiento. Se trabajó en el entorno Google Colab Pro, donde se cargó el conjunto de datos en la carpeta **content** del entorno, descomprimiendo los archivos para facilitar su manipulación. A continuación, se procesó el archivo de metadatos (*metadata.csv*), que contiene las referencias a los artículos científicos, y se seleccionaron únicamente los documentos almacenados en la carpeta **pmc_json**, correspondientes a artículos en formato XML.

Como parte de la limpieza del corpus, se eliminaron los artículos duplicados utilizando la información contenida en el archivo de metadatos, conservando únicamente las versiones más actualizadas. Tras este filtrado, se obtuvieron **373.674 artículos**, los cuales fueron divididos en dos subconjuntos: uno destinado al entrenamiento de

BioBERT y el otro al entrenamiento de SciBERT, en línea con la propuesta inicial de comparar ambos modelos.

Para preparar los textos, se aplicaron técnicas de **tokenización**, **etiquetado** y **normalización**, con el fin de convertir los artículos a un formato adecuado para el entrenamiento supervisado. Posteriormente, se construyeron tres conjuntos bien definidos: **entrenamiento**, **validación** y **prueba**. Además, se creó una nueva columna denominada **label**, que actúa como variable objetivo para realizar la **clasificación binaria** de los documentos (artículos sobre COVID-19 vs. no COVID-19).

Para la implementación del sistema R*etrieval-Augmented Generation* (RAG), se cargó una porción del contenido de la carpeta **pmc_json** en una ubicación específica dentro del entorno de Google Drive, con el objetivo de extraer el texto completo de los artículos y convertirlos en representaciones vectoriales. A partir de estos textos procesados, se construyó un índice de recuperación utilizando la librería FAISS (*Facebook AI Similarity Search*), generando así el archivo faiss_index_, que permite realizar búsquedas eficientes por similitud semántica entre las consultas y los documentos almacenados.

Este índice es fundamental para el correcto funcionamiento de RAG, ya que facilita la recuperación rápida de los documentos más relevantes antes de que el modelo generativo produzca la respuesta final.

Herramientas y entornos

- Python para la implementación y entrenamiento de los modelos.
- Hugging Face Transformers para el *fine-tuning* y tokenización de los modelos BioBERT entrenados.
- FAISS (*Facebook AI Similarity Search*) para la implementación de la parte de recuperación en RAG.
- PyTorch para la creación y entrenamiento de redes neuronales.

Hiperparámetros

- **BioBERT (Modelo Entrenado N° 3)**
  - Modelo base: BioBERT v1.1
  - Número de épocas: 3
  - Tasa de aprendizaje (*learning rate*): 2e-5
  - Decaimiento de pesos (*weight decay*): 0.01

- Warmup steps: 500
- Tamaño de lote (*batch size*):
  - Entrenamiento: 32 por dispositivo
  - Evaluación: 64 por dispositivo
- Evaluación:
  - Frecuencia: cada 500 pasos
- Métrica para el mejor modelo: pérdida de evaluación (*eval_loss*)
- Criterio de selección: menor pérdida (*greater_is_better* = *False*)
- Precisión mixta habilitada (fp16=True) para acelerar entrenamiento y reducir uso de memoria
- Scheduler lineal: para el ajuste de la tasa de aprendizaje (lr_scheduler_type="linear")
- Semilla fijada para reproducibilidad (seed): enviada mediante iteración.

- **RAG**:
  - Número de documentos recuperados: 25
  - Número de tokens generados: 13
  - Embedding model: OpenAIEmbeddings.
  - Chunk size: cada documento fue dividido en fragmentos de hasta 3.000 caracteres para mejorar la granularidad del índice.
  - Batch size: se procesaron los embeddings en lotes de 100 fragmentos, optimizando el uso de recursos computacionales.

Evaluación de resultados

Métricas

Para evaluar la efectividad de los modelos, se utilizan las siguientes métricas:

1. **Precisión (Accuracy)**: Mide el porcentaje de clasificaciones correctas en la clasificación binaria.
2. **Precisión y Recall**: En la evaluación de las respuestas generadas por RAG, se mide la precisión y el recall para evaluar la relevancia de las respuestas.

3. **F1-Score**: Se calcula para balancear la precisión y el recall, especialmente en tareas desbalanceadas.

Análisis cualitativo

Se llevó a cabo un análisis cualitativo de las respuestas generadas por el sistema RAG, con el objetivo de evaluar la coherencia, relevancia y precisión de la información recuperada. Para ello, se formularon preguntas específicas tanto en español como en inglés, permitiendo una evaluación visual de las respuestas y una comparación entre distintos idiomas. Entre las consultas realizadas se incluyeron: "¿Cuáles son los síntomas persistentes más comunes tras una infección por COVID-19?" y su equivalente en inglés, *"What are the most common long-term symptoms after COVID-19 infection?"*. Esta estrategia permitió verificar si el sistema era capaz de identificar y extraer información relevante de los artículos científicos, independientemente del idioma de entrada. Los resultados demostraron que el sistema ofreció respuestas consistentes y temáticamente equivalentes para ambas versiones de la pregunta, evidenciando su capacidad multilingüe en la recuperación de contenidos. Asimismo, se emplearon otras consultas en inglés, como *"Traditional Medicine Reveals Overuse as a Potential Risk for Aggravating COVID-19"*, con el fin de explorar si el sistema era capaz de localizar documentos que contuvieran evidencia empírica o discusión directa sobre el tema consultado. En todos los casos, las respuestas generadas permitieron confirmar que el sistema RAG con el modelo de clasificación binaria BioBERT con *fine-tuning*, puede recuperar información relevante y coherente con las preguntas planteadas, incluso cuando estas se formulan en distintos idiomas.

Recomendaciones

- Se sugiere ampliar la base documental multilingüe para el sistema RAG, a modo de poder fortalecer la capacidad del sistema en otros idiomas distintos al inglés y al español.
- Es recomendable integrar una evaluación cuantitativa complementaria que permita medir precisión, cobertura y redundancia de las respuestas, además del análisis cualitativo.
- Se propone incorporar un módulo de retroalimentación por parte del usuario para mejorar la personalización y la pertinencia de las respuestas generadas. Parecido a la selección de respuestas planteadas por ChatGPT.
- Para aplicaciones clínicas, sería conveniente validar las respuestas del sistema frente a criterios expertos del dominio médico.

Conclusiones

El sistema RAG con el modelo de clasificación binaria BioBERT entrenado demostró un desempeño sólido en la recuperación de información relevante, precisa y coherente, tanto en español como en inglés. La comparación entre preguntas semánticamente equivalentes formuladas en diferentes idiomas evidenció su capacidad para ofrecer respuestas consistentes, lo que respalda su aplicabilidad en contextos multilingües. La calidad de las respuestas sugiere que este tipo de sistemas puede ser una herramienta útil en entornos de investigación biomédica, particularmente en escenarios que requieren análisis rápido de literatura científica extensa.

Cabe destacar que el sistema fue respaldado por el entrenamiento de tres variantes del modelo BioBERT mediante *fine-tuning*, lo que permitió explorar distintas parametrizaciones y configuraciones de entrenamiento. Este enfoque experimental facilitó la identificación del modelo con mejor rendimiento en tareas de clasificación binaria sobre documentos científicos del corpus CORD-19. La evaluación comparativa entre modelos no solo mejoró la precisión del sistema, sino que también optimizó su capacidad de generalización frente a nuevas consultas.

Si bien la propuesta inicial contemplaba la experimentación con el modelo SciBERT, no fue posible llevar a cabo su entrenamiento en esta fase del trabajo. No obstante, se considera altamente recomendable realizar dicha experimentación en investigaciones futuras, dado el enfoque de SciBERT en literatura científica y su potencial para complementar o incluso superar el rendimiento de BioBERT en ciertos dominios.

Futuras mejoras podrían enfocarse en refinar la cobertura temática, la precisión contextual y la adaptabilidad a diferentes dominios de conocimiento. Adicionalmente, la incorporación de mecanismos de retroalimentación y validación humana permitiría robustecer aún más la utilidad práctica del sistema en entornos reales.

Referencias Bibliográficas

1. **Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J.** (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
2. **Yoon, W., So, C. H., Lee, J., & Kang, J.** (2021). Pre-trained language model for biomedical named entity recognition. *Briefings in Bioinformatics, 22*(6), bbab228. https://doi.org/10.1093/bib/bbab228

3. **Beltagy, I., Lo, K., & Cohan, A.** (2019).
SciBERT: A pretrained language model for scientific text. *Proceedings of EMNLP 2019*. https://arxiv.org/abs/1903.10676

4. **Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., ... & Wang, K.** (2020).
CORD-19: The COVID-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706*. https://arxiv.org/abs/2004.10706