Jaime Rueda 1566320

INFORME DE PROGRESO/PROPUESTA FINAL

Título del trabajo: AI-based journal editor assessment of scientific research articles

Contexto de la problemática: En el panorama actual del mundo de las publicaciones científicas, nos encontramos con varios problemas a la hora de publicar un artículo nuevo y que sea aceptado por la comunidad. Para empezar, se necesita pagar una suma de alrededor de 250€, conocido como Article Processing Charges para su publicación, lo cual no es necesariamente un problema puesto que para publicar y mantener en la red estos artículos se tiene un coste obligado asociado. Ahora bien, el problema principal surge cuando el autor de dicho artículo pretende publicar en una revista de alta categoría donde independientemente de la calidad del trabajo se cargan sumas más de 10 veces mayores. Además, entran en juego otros peligros como el rechazo a la publicación, ser pionero en un campo específico o la demora en obtener una resolución.

El impacto de un artículo se define mayoritariamente con el paso del tiempo, según factores como las citas o referencias que se le haga. Pero por otro lado, la categoría de la revista en la que se haya publicado puede influenciar enormemente la validación de la comunidad científica. La alternativa de publicar en otras plataformas como Diamond Open Access, cuyos costes son mucho menores, no tienen el mismo valor pues se enfrentan a diversos problemas como la ausencia de editores que revisen el trabajo, la jerarquía según la calidad del trabajo o el prestigio de la propia plataforma en la comunidad.

El objetivo de este TFG es buscar una posible solución a toda esta serie de problemas apoyándose en la Inteligencia Artificial. Al mismo tiempo, crear y entrenar un modelo desde cero sería una cantidad de trabajo que sobrepasa las capacidades del proyecto, por lo que es más adecuado aprovechar modelos de lenguaje grande “LLM” existentes y ya entrenados y altamente complejos. Para así poder predecir el posible impacto de un artículo científico en el momento de su publicación según el abstract.

Objetivo general: Diseñar un sistema basado en el modelo de Large Language Models (GPT-4o) que simule las funciones de validación de los editores de revistas científicas.

Objetivo específico: Reentrenar a GPT-4o mediante el *prompting* adecuado para que sea capaz de predecir el impacto que tiene un artículo según su abstract.

METODOLOGIA

1. Diseño general del sistema

La utilización de GPT-4o se basa en su alta capacidad para manejar frases y textos complejos sobre temas específicos en comparación con sus competidores como Mistral o Llama.

El diseño general del sistema se basa en los siguientes componentes:

- *Elaboración del preprompt*: incluye un *training set* con información sobre los artículos científicos de entre 2020 y 2023, una *query* basada en *promp engineering* explicando la función del LLM y *reviews* indicando el contexto del campo de investigación.
- *Elaboración del prompts*: incluye el *dataset* con la información de los artículos del año 2024 sobre los que se realizará la validación.
- *Análisis de los resultados*: comprobación y comparación de los resultados obtenidos.

2. Recolección y preparación de datos

Para la consecución del proyecto no se requiere del cuerpo completo del *paper,* únicamente con el abstract y la referencia bibliográfica se pretende analizar el impacto. Estos datos son de orden público por lo que se han obtenido de páginas web como Pubmed o Semantic Scholar.

Los criterios de selección de los artículos son los relacionados con la temática de *bacteria transcription regulator represor activator promoter*. La elección de este campo se justifica con que es altamente específico y sobre el que se pretende que la LLM adquiera conocimientos de alto nivel.

El procesamiento de datos ha consistido en su obtención mediante llamadas a la API de Pubmed y una API privada que fue proporcionada por Semantic Scholar. Estos datos se reconviertieron en ficheros .txt y reducidos a la información principal. Posteriormente se utilizaron para el *preprompt* y los *prompts*.

Para medir el impacto de los diferentes *papers* se han probado diferentes métricas como las citas o el factor SJR de Scimago. Este factor es un indicador de impacto de revistas científicas que se calcula utilizando datos de Scopus y que pondera las citas en función de la reputación de la revista que la realiza (1).

3. Diseño del prompting

El diseño del prompting se basa en el artículo científico "*The prompt report*" (2). Para la elaboración del *preprompt* se otorga a la LLM un rol de editor de artículos con sus respectivas características como el tono, el estilo o la profesionalidad. En el siguiente paso se le concede una breve guía para ayudar a valorar los artículos y se le adjuntan dos archivos .txt de los cuales se explica todo su formato. El primer fichero .txt contiene el *training set* que incluye la siguiente información de los artículos científicos entre los años 2020 y 2023:

- Title
- Publisher
- Abstract
- Year
- SJR Quartile*

**La utilización del SJR quartile se justifica con que las LLM funcionan mejor clasificando de manera categórica que proponiendo un número, ya que no funcionan bien con la aritmética. Por esta razón no se ha usado como métrica de impacto la cantidad de citas o el valor de SJR.*

El segundo fichero txt. es una *review* sobre los campos de investigación que otorgan contexto al editor.

Para la elaboración de los *prompts* se utilizan una serie de artículos del año 2024 sobre la misma temática incluida en el *training set* y se pide a la LLM que trate de predecir a que cuartil pertenece la revista en la que se debería publicar cada artículo.

La comunicación con el LLM GPT-4o se realiza mediante el uso de la API. Se mide el tamaño de las *queries* para que el número de *tokens* sea menor que la *Window size* del LLM (128.000 *tokens*).

4. Evaluación del sistema

La evaluación del sistema se basa en los siguientes criterios de evaluación:

- Precisión: porcentaje de artículos cuyo cuartil predicho por GPT-4o coincide con el cuartil real de la revista en la que fue publicado.
- Distribución de errores: analizar si los errores son sistemáticos (sobreestimación hacia cuartiles altos) o aleatorios.

Para realizar la evaluación se utilizará el set de validación con artículos científicos del año 2024 sobre el mismo campo temático (*bacterial transcription regulation*) y que no formaban parte del conjunto de entrenamiento. A cada artículo se le aplicará el *prompt* desarrollado, y se registrará la predicción de cuartil realizada por GPT-4o. posteriormente se contrastará la predicción del modelo con el cuartil SJR real de la revista donde se publicó el artículo, utilizando datos obtenidos de Scopus o Scimago Journal Rankings.

CAMBIOS EN EL DESARROLLO DEL PROYECTO

1. Replanteamiento de los objetivos

Inicialmente el objetivo del proyecto era reentrenar a GPT-4o mediante el *prompting* adecuado para que fuera capaz de predecir el impacto que tiene un artículo según su abstract. Sin embargo, no hay una métrica establecida ni validada a nivel global que mida este impacto. Solo son métricas de valor arbitrario como el autor, las citas, el prestigio de la revista, el campo de investigación, etc. Para poder focalizar el trabajo en un objetivo real y con resultados palpables, se ha decidido utilizar el ranking elaborado por la comunidad científica de Scimago.

Finalmente, se ha replanteado el **objetivo específico**: Reentrenar a GPT-4O mediante el *prompting* adecuado para que prediga la categoría de la revista en la cual debería ser publicado.

2. Replanteamiento de las métricas de evaluación

Se consideraron las citas como métricas de evaluación incluyendo las citas "*highly influential*" que ofrece la página de Scemantic Scholar. Éstas son útiles dado el posible problema de que un artículo solo hubiera sido citado una o pocas veces, pero fuera fundamental para otros artículos de gran relevancia. Se descartó esta métrica porque los modelos LLM como GTP-4o no trabajan bien con modelos numéricos y haría una predicción prácticamente infundada.

Por otro lado, utilizar varios LLM diferentes fue descartado porque para hacer el *prompting* en condiciones adecuadas se necesita un gran tamaño de *window size.* Los LLM gratuitos son formatos con ventanas muy pequeñas, con lo que se ha preferido trabajar con un LLM de pago para mayor *window size*.

3. Reconstrucción de los cuartiles

Analizando *datasets* sobre temas muy específicos, como es el caso que se ha seleccionado para este trabajo, se ha comprobado que la separación de cuartiles proporcionada por Scimago es poco discriminatoria porque la mayoría de las revistas se encuentran en el Q1 (el ranking de más prestigio). Para balancear los datos y no crear un sesgo, se ha procedido a hacer una nueva categorización de los cuartiles en función del valor SJR.

BIBLIOGRAFIA:

1. *SJR : Scientific Journal Rankings*. (s. f.). https://www.scimagojr.com/journalrank.php

2. *The Prompt report*. (s. f.). https://sanderschulhoff.com/Prompt_Survey_Site/