Proyecto Redes de Computadores - Herramienta para Investigaciones Académicas

1st Luis David Preciado Martínez
Ingeniería Mecatrónica
Universidad Nacional de Colombia
Bogotá, Colombia
lpreciadom@unal.edu.co

I. Resumen

Este informe presentará el desarrollo de una herramienta para la evaluación y selección de fuentes bibliográficas para los investigadores académicos. Ya que actualmente estos se ven enfrentados ante una gran variedad de opciones de publicaciones científicas de diversas calidades. Por lo cual la herramienta permitirá averiguar la calidad de las revistas indexadas tanto colombianas como internacionales y además implementar el análisis de textos de ChatGPT para hacer preguntas sobre las fuentes seleccionadas.

II. Abstract

This report will present the development of a tool for the evaluation and selection of bibliographic sources for academic researchers. Since these are currently faced with a wide variety of options for scientific publications of different qualities. Therefore, the tool will allow to find out the quality of both Colombian and international sources and also implement ChatGPT text analysis to ask questions about the selected sources.

***Palabras claves*—Fuentes bibliográficas, revistas indexadas, ChatGPT, inteligencia artificial**

III. Introducción

En el contexto de la investigación científica, la calidad y confiabilidad de las fuentes utilizadas desempeñan un papel fundamental en la generación de conocimiento sólido y válido. La relevancia y credibilidad de los resultados dependen en gran medida de la selección cuidadosa de fuentes respaldadas por rigurosos procesos de revisión y verificación. En este informe, se abordará la importancia crucial de emplear fuentes de alta calidad en el proceso de investigación, destacando cómo una base sólida de literatura científica influye en la objetividad, la precisión y la solidez de los hallazgos. Además, se explorarán los criterios clave para evaluar la calidad de las fuentes, junto con los riesgos asociados con el uso de información no confiable. Es relevante señalar que en el ámbito académico, herramientas de evaluación y clasificación de revistas y conferencias, como Scimago Journal & Country Rank, así como sistemas de indexación nacionales como Publindex en Colombia, proporcionan orientación valiosa al investigador al identificar fuentes respetadas y de alto impacto en sus respectivas disciplinas. En última instancia, este informe busca subrayar la necesidad de una selección informada y crítica de fuentes, respaldando así la integridad y excelencia en el ámbito de la investigación científica.

Las revistas predatorias, a menudo descritas como un flagelo en el ámbito de la literatura científica, representan una amenaza sustancial para la integridad de la investigación y la comunicación académica. Estas revistas explotan a los autores al cobrar tarifas de publicación sin proporcionar una revisión por pares adecuada, lo que resulta en la publicación de artículos deficientes que malgastan los esfuerzos y recursos de los investigadores. En este sentido, los estudios que examinan las dinámicas detrás de las presentaciones de los autores a tales revistas han brindado perspicaces ideas [1]. Los autores, especialmente aquellos de países de ingresos bajos y medianos (LMIC), pueden contribuir de manera consciente o inconsciente a revistas predatorias, percibiéndolas como un medio para navegar los desafíos de un entorno académico competitivo [1]. Los investigadores de LMIC a menudo enfrentan dificultades para comprender las complejidades de la publicación científica moderna, exacerbadas por barreras idiomáticas y un conocimiento limitado de las normas de publicación [1]. Sorprendentemente, muchos autores no son conscientes de que sus envíos son objeto de interés por parte de editores predatorios, lo que ilustra la magnitud del desafío. Algunos incluso se involucran erróneamente con invitaciones a encuestas, adjuntando artículos para solicitar una "publicación rápida.º preguntar sobre costos [1].

En el ámbito de la ciencia médica, el surgimiento de compañías de revistas predatorias constituye una preocupación alarmante. Estas entidades capitalizan las aspiraciones de los autores a través de campañas de marketing agresivas y altas tasas de aceptación, solo para descuidar la revisión por pares adecuada y poner en peligro la solidez científica del trabajo publicado [5]. Dado que pacientes y audiencias más amplias confían en la información difundida por la literatura científica, la proliferación de tales revistas socava la credibilidad de la ciencia médica, lo que requiere conciencia entre científicos, médicos e incluso pacientes [5]. Reconociendo esto, ha surgido la imperiosa necesidad de informar a autores jóvenes e inexpertos sobre el daño potencial a su reputación, las deficiencias de los procesos de revisión por pares inadecuados y el riesgo de que cierren revistas enteras, lo que podría resultar en la pérdida de contenido publicado [5].

En medio de estas preocupaciones, la denominación de

revistas como "predatorias"debido a prácticas deshonestas de revisión por pares y costos de publicación se ha vuelto frecuente en el panorama del acceso abierto [4]. Un examen exhaustivo de los perfiles de los autores en distintos grupos de revistas ha iluminado tendencias relevantes. Los autores que contribuyen a revistas "predatorias"son principalmente jóvenes investigadores de naciones en desarrollo, un patrón que se cree está influenciado por los contextos económicos y socioculturales prevalentes en LMIC [4]. Esto subraya la necesidad de esfuerzos enfocados para educar a los investigadores sobre los riesgos de la publicación predatoria y para cerrar las brechas en la comprensión del panorama más amplio de la publicación académica [4]. A la luz de estos hallazgos, la salvaguardia de la calidad y autenticidad de la literatura científica requiere una acción colectiva entre instituciones e investigadores en todo el mundo.

La aparición de tecnologías avanzadas de inteligencia artificial, como ChatGPT, ha inaugurado una era transformadora en diversos ámbitos, incluyendo la educación y la investigación [1,2]. ChatGPT, un notable modelo de lenguaje desarrollado por OpenAI, tiene el potencial de influir significativamente en la forma en que se difunde la información y se adquiere el conocimiento. Las capacidades de esta herramienta se han explorado en una multitud de contextos, con investigadores considerando tanto sus beneficios como sus desafíos.

En la publicación científica, la introducción de ChatGPT plantea preguntas intrigantes sobre su impacto en el panorama de la investigación [3]. Como se demostró en el estudio de caso de Manohar y Prasad (2023), el software puede generar texto coherente y gramaticalmente correcto, sirviendo como una fuente potencial de contenido introductorio [3]. Es importante destacar que la capacidad de la herramienta para producir rápidamente cantidades sustanciales de texto ofrece ventajas en términos de eficiencia [3]. Sin embargo, surgen preocupaciones con respecto a la profundidad y singularidad del contenido generado. El texto producido a menudo carece de una "voz"distintiva, asemejándose a la escritura científica convencional caracterizada por su naturaleza seca y formulaica [3]. Esto lleva a considerar si la tecnología fomentará inadvertidamente la conformidad en los estilos de escritura o si los investigadores diversificarán conscientemente su expresión.

No obstante, las limitaciones de ChatGPT se hacen evidentes al considerar los requisitos multifacéticos de la escritura científica [3]. Su texto carece de la profundidad y la evaluación crítica esencial para el empeño científico, especialmente en campos en rápida evolución como la neurociencia [3]. Si bien el software puede agregar y presentar información de diversas fuentes, carece de la interpretación matizada, la comprensión del contexto y la capacidad para vincular ideas complejas que definen el pensamiento crítico de alto nivel en la ciencia [3]. Además, la opacidad de sus fuentes de datos y los procesos de toma de decisiones plantea preocupaciones sobre la precisión y la confiabilidad de la información que genera [3]. A diferencia de los investigadores humanos, la herramienta no puede ser considerada responsable por sus opiniones o decisiones, lo que la hace inadecuada para el discurso académico en el que el rigor intelectual y la responsabilidad son primordiales.

En el ámbito educativo, ChatGPT presenta tanto oportunidades como desafíos. Los educadores tienen formas novedosas de interactuar con los estudiantes a través de conversaciones interactivas, retroalimentación personalizada e innovadores métodos de enseñanza [2]. Sin embargo, surgen preocupaciones sobre el posible engaño durante los exámenes en línea, el riesgo de disminuir las habilidades de pensamiento crítico debido a una dependencia excesiva de la herramienta y el desafío de evaluar el trabajo generado por ChatGPT [2]. En el campo de la educación en programación, se ha explorado el papel de ChatGPT en mejorar las habilidades de programación de los estudiantes a través de la generación de código, la creación de seudocódigo y la corrección de código [2]. Si bien la tecnología demuestra promesa en el apoyo al aprendizaje de programación, sus limitaciones en la replicación de la creatividad humana y el pensamiento analítico siguen siendo evidentes [2].

Dicho este contexto de la problemática se presentan los siguientes objetivos para este proyecto:

- Poder procesar la información contenida en un documento de tipo académico por medio de ChatGPT.
- Ofrecer una evaluación de la fuente de dicho documento en términos de la calidad de la revista que hizo la publicación.
- Poder realizar preguntas a ChatGPT sobre el contenido del documento y obtener respuestas a estas.

IV. Trabajos Relacionados

Actualmente se cuentan de manera abierta con las bases de datos de Scimago (International Scientific Journal & Country Ranking) y Publindex del Ministerio de Ciencias de Colombia. Pero no se cuenta con una herramienta de fácil acceso a ambas bases de datos, y aun menos incluyendo funcionalidades de ChatGPT. Scimago tiene un buscador bastante competente, pero Publindex no.

Publindex es un índice bibliográfico colombiano que categoriza y clasifica las revistas científicas publicadas en Colombia. La clasificación proporcionada por Publindex indica la calidad y el impacto de estas revistas dentro de la comunidad académica e investigadora del país. Las clasificaciones se basan en varios criterios, incluyendo la calidad editorial de la revista, el proceso de revisión por pares, la frecuencia de publicación y las métricas de citas. Existen diferentes niveles o categorías en la clasificación de Publindex, cada uno representando diferentes grados de calidad e impacto. Estas categorías incluyen:

1. Publindex A1: Las revistas en esta categoría se consideran de la más alta calidad e impacto. Por lo general, tienen procesos rigurosos de revisión por pares, altos estándares editoriales y una influencia significativa en su campo.
2. Publindex A2: Las revistas en esta categoría también tienen alta calidad e impacto, aunque podrían tener ligeramente menos citas o un proceso de revisión por pares un poco menos estricto en comparación con las revistas A1.

3. Publindex B: Las revistas en esta categoría tienen buena calidad e impacto, pero podrían no cumplir con los criterios para las revistas A1 o A2. Aún se considera que son respetables y contribuyen de manera significativa a sus respectivos campos.
4. Publindex C: Las revistas en esta categoría podrían tener un enfoque más regional o especializado y podrían no tener un impacto tan amplio. Sin embargo, siguen contribuyendo al discurso académico dentro de su nicho.
5. Publindex D: Las revistas en esta categoría podrían ser más nuevas o tener un impacto más limitado. Aún se reconoce que son fuentes legítimas de información académica.

Es importante tener en cuenta que las clasificaciones de Publindex son específicas de Colombia y se utilizan para evaluar la calidad de las revistas científicas colombianas. Los investigadores y académicos dentro de Colombia suelen referirse a estas clasificaciones al evaluar la credibilidad e impacto de las revistas para su trabajo. Las clasificaciones también pueden influir en decisiones relacionadas con la financiación y el reconocimiento académico. Sin embargo, fuera de Colombia, es posible que estas clasificaciones no tengan el mismo nivel de reconocimiento o influencia.

El Scimago Journal & Country Rank (SJR) es un portal público que proporciona varias métricas e indicadores para revistas científicas y países. Una de las características del SJR es su sistema de clasificación por cuartiles, que se utiliza para categorizar revistas en función de su impacto dentro de sus respectivos campos. El sistema de clasificación por cuartiles divide las revistas en cuatro cuartiles, cada uno representando un segmento de la distribución de revistas dentro de una categoría temática específica. Estos cuartiles se calculan en función del indicador SJR, que tiene en cuenta no solo el número de citas recibidas por una revista, sino también el prestigio de las revistas que citan. Esto es lo que representa cada cuartil:

1. Q1 (Top 25 %): Las revistas en el primer cuartil se consideran las más prestigiosas e impactantes dentro de su categoría temática. Están entre el 25 % superior de las revistas en términos de valores del indicador SJR. Estas revistas suelen tener un alto número de citas y se consideran influyentes en su campo.
2. Q2 (Segundo 25 %): Las revistas en el segundo cuartil aún son bien consideradas y tienen un impacto significativo, pero no son tan influyentes como las del primer cuartil. Se sitúan en el rango del 25 al 50 % en términos de valores del indicador SJR.
3. Q3 (Tercer 25 %): Las revistas en el tercer cuartil se consideran que tienen un impacto moderado dentro de su categoría temática. Se sitúan en el rango del 50 al 75 % en términos de valores del indicador SJR. Si bien estas revistas pueden no tener un impacto tan alto como las de los cuartiles superiores, siguen contribuyendo a la literatura académica.
4. Q4 (Último 25 %): Las revistas en el cuarto cuartil tienen el impacto más bajo dentro de su categoría temática. Se encuentran entre las revistas menos citadas y menos influyentes, situándose en el 25 % inferior en términos de valores del indicador SJR.

Es importante tener en cuenta que la clasificación por cuartiles es relativa a la categoría temática. Una revista que esté en el Q1 en una categoría temática podría estar en un cuartil diferente en otra categoría temática. Además, el sistema de clasificación por cuartiles es solo una forma de evaluar el impacto y el prestigio de una revista. Los investigadores suelen considerar varias métricas e indicadores al evaluar la importancia de una revista dentro de su campo.

Predatory Reports es una organización compuesta por investigadores voluntarios que han experimentado perjuicios a manos de editoriales depredadoras y han decidido trabajar de manera anónima para proporcionar una valiosa ayuda a la comunidad académica. Su misión principal es asistir a otros investigadores en la identificación de revistas y editoriales de confianza al recopilar y difundir información pública sobre prácticas cuestionables en el ámbito de la publicación científica. Lo que distingue a Predatory Reports es su compromiso de ofrecer todos sus servicios de forma totalmente gratuita, sin mostrar publicidad en su sitio web ni recibir apoyo financiero de ninguna empresa, asumiendo los costos por sí mismos. Este enfoque sin ánimo de lucro demuestra su dedicación a la promoción de la integridad en la investigación científica y a la construcción de confianza en las publicaciones académicas. Además, la organización enfatiza que no se basa en la autoridad de las fuentes, sino que recopila información pública para ayudar a los autores a tomar decisiones informadas sobre dónde publicar sus investigaciones. Cada una de sus publicaciones incluye una lista de referencias para que los usuarios puedan verificar la información por sí mismos, fomentando así la transparencia y la autonomía de la comunidad académica. Predatory Reports también está abierta a colaboraciones y sugerencias, lo que refleja su compromiso con la mejora continua y su disposición a trabajar en conjunto con otros interesados en la promoción de prácticas éticas en la publicación científica. A medida que su presencia en Internet crece, la organización reconoce que enfrenta amenazas, lo que subraya la importancia de su trabajo en la denuncia de prácticas depredadoras en la industria editorial.

V. Metodología

La metodología del proyecto se divide en varias fases, cada una con sus tareas específicas:

V-A. Diseño de la Interfaz de Usuario

Esta fase se centra en la creación de una interfaz de usuario intuitiva y amigable utilizando la biblioteca Tkinter de Python. La interfaz incluye elementos como un menú desplegable de selección de país, un campo de entrada de texto para el nombre de la revista y botones para cargar documentos PDF y realizar preguntas a un chatbot. Se proporcionan etiquetas claras e instrucciones para guiar a los usuarios.

V-B. Obtención de Datos

En esta etapa, se utilizan conjuntos de datos en formato CSV, como `publindex.csv`, `scimagojr 2022.csv` y `predatory.csv`, que contienen información sobre revistas colombianas, revistas internacionales y una lista de revistas depredadoras. La biblioteca Pandas se utiliza para leer y transformar estos datos en dataframes para un manejo eficiente.

V-C. Búsqueda y Clasificación de Revistas

La funcionalidad central de la aplicación consiste en buscar revistas y proporcionar información sobre su clasificación. El usuario ingresa el nombre de la revista, y se realiza una búsqueda en función del país seleccionado. Se intenta una coincidencia exacta, y si no se encuentra, se aplica una estrategia de coincidencia parcial. La información relevante se extrae y muestra al usuario.

V-D. Presentación de Datos

Una vez completada la búsqueda, los resultados se muestran en un diálogo personalizado. Este diálogo incluye información como la categoría de la revista y la información de clasificación de la institución. Además, se proporciona un botón para realizar una segunda búsqueda en las bases de datos y verificar si la revista se encuentra reportada por The Predatory Reports.

V-E. Q&A con el Documento PDF

Esta fase implica la integración de los APIs de OpenAI y LlamaAI para cargar documentos PDF y realizar preguntas basadas en su contenido. Se utiliza un lector de directorios para cargar documentos desde una carpeta específica, se genera un índice global para búsquedas futuras y se permite a los usuarios seleccionar archivos PDF para su análisis. La función de preguntas y respuestas facilita la interacción del usuario con el chatbot y presenta respuestas basadas en el contenido del documento.

VI. Diseño e Implementación

VI-A. Diseño de la Interfaz de Usuario

La fase de diseño de la interfaz de usuario se centra en crear una plataforma intuitiva y amigable para interactuar con el sistema de clasificación de revistas. Se utiliza Tkinter, una biblioteca de Python para el desarrollo de interfaces gráficas, para diseñar e implementar los elementos gráficos. La interfaz de usuario consta de una ventana principal con un título, instrucciones, un menú desplegable de selección de país, un campo de entrada de texto para el nombre de la revista y un botón "Buscar". Se proporcionan etiquetas claras e instrucciones para guiar a los usuarios en la introducción de datos y comprender el propósito de la aplicación.

Además, se cuenta con botones que permiten cargar un documento en formato PDF, analizar su contenido y realizar una pregunta a ChatGPT basado en su contenido. Este proceso se explicará en mayor detalle en una siguiente sección.

VI-B. Obtención de Datos

Para facilitar el sistema de clasificación, se utilizan tres conjuntos de datos. El conjunto de datos `publindex.csv` contiene información sobre revistas colombianas y sus clasificaciones, mientras que el conjunto de datos `scimagojr 2022.csv` contiene datos sobre revistas internacionales. Además el conjunto `predatory.csv` contiene una lista de revistas consideradas depredadoras publicada por Predatory Reports. Se emplea la biblioteca Pandas para leer estos archivos CSV y transformarlos en dataframes. Este proceso garantiza un almacenamiento y manipulación eficiente de los datos para los pasos siguientes.

VI-C. Búsqueda y Clasificación de Revistas

La funcionalidad central de la aplicación radica en su capacidad para buscar revistas y proporcionar sus clasificaciones. El usuario ingresa el nombre de la revista que desea buscar, y esta entrada se almacena en la variable `search_name`. En función del país seleccionado, se elige el conjunto de datos de Publindex o Scimago para un procesamiento adicional.

Primero se intenta una coincidencia exacta convirtiendo tanto el nombre de búsqueda como los valores del conjunto de datos a minúsculas y eliminando acentos. Si se encuentra una coincidencia exacta, se extrae y almacena la información relevante, como la categoría y el nombre de la revista.

En ausencia de una coincidencia exacta, se aplica una estrategia de coincidencia parcial. Los nombres de las revistas se despojan de espacios y acentos, y se realiza una comparación. Si se encuentran coincidencias parciales, se muestra al usuario la información relevante. Si se encuentran múltiples coincidencias parciales, se muestra un mensaje que insta al usuario a proporcionar una consulta más específica.

VI-D. Presentación de Datos

Una vez completado el proceso de búsqueda, la aplicación muestra el resultado en un diálogo de información personalizado. Este diálogo incluye una etiqueta de título que muestra el estado del resultado, un widget de texto para mostrar un mensaje detallado y un botón “Aceptar“ para cerrar el diálogo. La información mostrada incluye la categoría de la revista y la información de clasificación específica de la institución. Además de el botón de aceptar, se encuentra un botón de “Análisis de Revistas Depredadoras“ que permite realizar una segunda búsqueda en las bases de datos para revisar si la revista que se esta buscando, se encuentra reportada por The Predatory Reports.

VI-E. Q&A con el Documento PDF

Usando los APIs de OpenAI y LlamaAI se programa la funcionalidad de cargar un documento PDF y realizar preguntas basadas en el contenido de este. En la función `analyze_documents()`, los documentos se cargan desde un directorio utilizando un `SimpleDirectoryReader`. Este lector de directorios tiene opciones para la búsqueda

recursiva y la exclusión de archivos ocultos, lo que asegura un proceso completo de recuperación de documentos. Posteriormente, genera un índice global, para futuras operaciones de búsqueda y recuperación. La función `load_and_save_pdf()` permite a los usuarios seleccionar archivos PDF para su análisis mediante la apertura de un cuadro de diálogo de selección de archivos. Los archivos PDF seleccionados se guardan en una carpeta de destino designada. Además, esta función también maneja casos en los que es necesario crear la carpeta de destino o eliminar archivos existentes, proporcionando un lienzo en blanco para nuevas adiciones de documentos. Por último, la función `preguntar_button_click()` facilita la interacción del usuario con un chatbot o sistema de preguntas y respuestas. Obtiene preguntas ingresadas por el usuario, consulta el índice de documentos en busca de información relevante y presenta la respuesta.

VII. Resultados y Pruebas

VII-A. Diseño de la Interfaz de Usuario

Se construyó una interfaz de usuario amigable que permita acceder a todas las funcionalidades del programa.

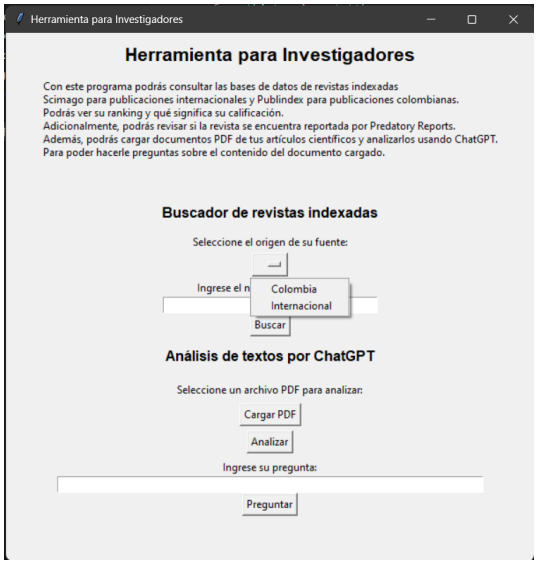

Figura 1. Ventana Inicial

VII-B. Búsqueda y Clasificación de Revistas

El programa cuenta con la capacidad de realizar búsquedas en `publindex.csv`, `scimagojr 2022.csv` y `predatory.csv`, que contienen información sobre revistas colombianas, revistas internacionales y una lista de revistas depredadoras.

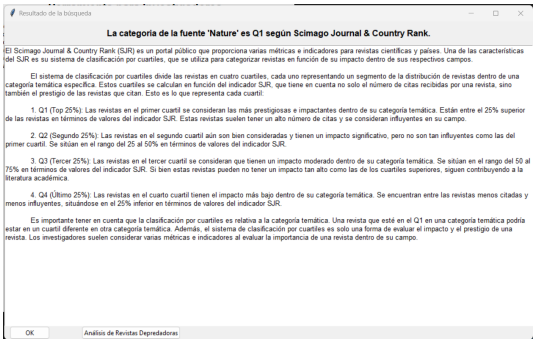

Figura 2. Ventana Resultados de Búsqueda

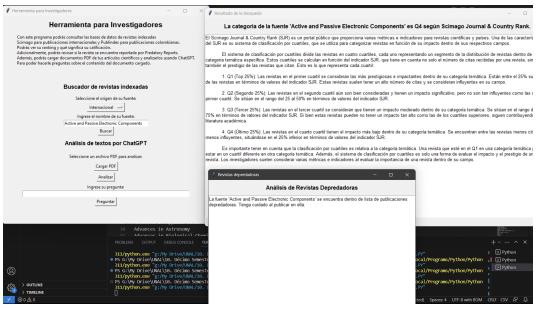

Figura 3. Búsqueda Revista Depredadora

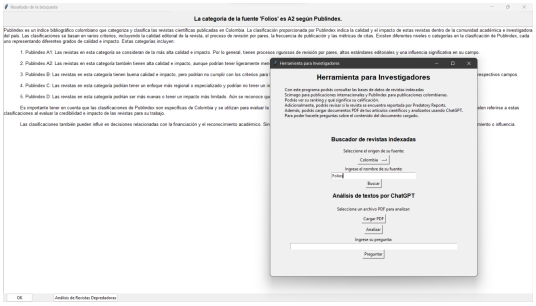

Figura 4. Búsqueda Revista Colombiana

VII-C. Q&A con el documento PDF

Se implemento la funcionalidad de carga y análisis de documentos PDF, para posteriormete hacer preguntas sobre su contenido.

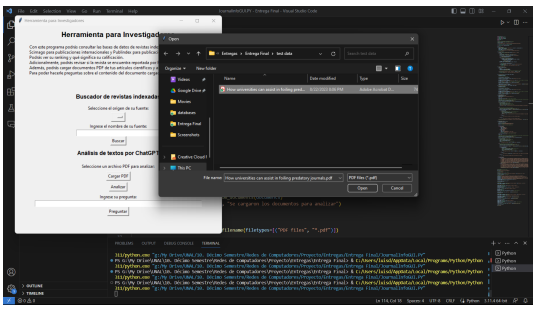

Figura 5. Carga de PDF

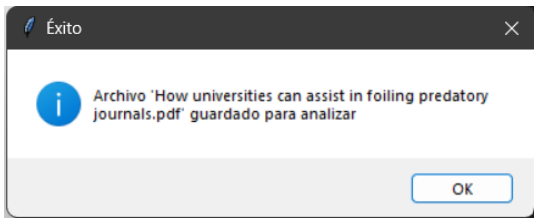

Figura 6. Éxito al cargar

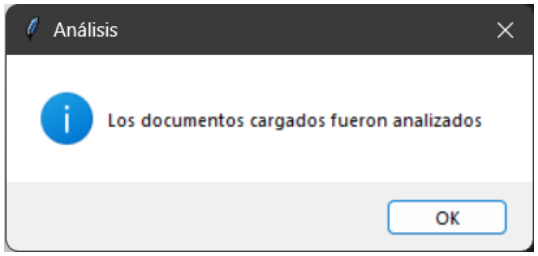

Figura 7. Análisis del PDF

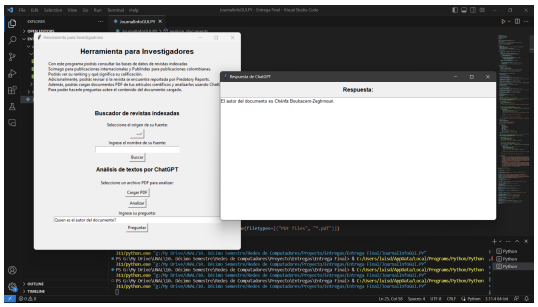

Figura 8. Pregunta y respuesta

VIII. Conclusiones

En conclusión, se logró el objetivo de desarrollar una herramienta para investigadores que permita evaluar la calidad de la fuente que se quiere utilizar o publicar. Además de brindar la posibilidad de cargar el documento que se está estudiando y hacer preguntas basadas en su contenido. Como aspectos para mejorar en versiones futuras, se podrían implementar más métricas de calificación de las revistas, como cantidad de publicaciones anuales o tiempos de revisión. Esto se intentó implementar en esta primera versión del programa, pero no se encontraron bases de datos que tuvieran esta información para los 3 grupos de revistas (nacionales, internacionales y depredadoras). Por lo cual, en vez de dar una implementación a medias, se dejó como un reto para versiones posteriores.

IX. Referenncias

[1] C. Boukacem-Zeghmouri, "Predatory journals entrap unsuspecting scientists. here's how universities can support researchers," Nature, vol. 620, no. 7974, pp. 469–469, 2023. doi:10.1038/d41586-023-02553-1

[2] Md. M. Rahman and Y. Watanobe, "Chatgpt for Education and research: Opportunities, threats, and strategies," Applied Sciences, vol. 13, no. 9, p. 5783, 2023. doi:10.3390/app13095783

[3] E. L. Hill-Yardin, M. R. Hutchinson, R. Laycock, and S. J. Spencer, "A chat(gpt) about the future of Scientific Publishing," Brain, Behavior, and Immunity, vol. 110, pp. 152–154, May 2023. doi:10.1016/j.bbi.2023.02.022

[4] J. Xia et al., "Who publishes in 'predatory' journals?," Journal of the Association for Information Science and Technology, vol. 66, no. 7, pp. 1406–1417, 2014. doi:10.1002/asi.23265

[5] G. Richtig, M. Berger, B. Lange-Asschenfeldt, W. Aberer, and E. Richtig, "Problems and challenges of predatory journals," Journal of the European Academy of Dermatology and Venereology, vol. 32, no. 9, pp. 1441–1449, 2018. doi:10.1111/jdv.15039