Facultad de Administración y Dirección de Empresas

Trabajo de Fin de Grado

Metaestudio de desinformación online relacionada con la salud

Ana Laura Correa Ribeiro

Julio 2025

Trabajo de Fin de Grado presentado en la Facultad de Administración y Dirección de Empresas de la Universidad de Santiago de Compostela para la obtención del Grado en Empresa y Tecnología

Resumen

En los últimos años, el uso de buscadores y redes sociales para consultar dudas sobre salud se ha convertido en una práctica habitual. El auge de internet ha facilitado el acceso rápido a la información, pero también ha incrementado el riesgo de exposición a contenidos falsos y dañinos. La desinformación médica no solo genera confusión, sino que puede derivar en la toma de decisiones erróneas, con consecuencias graves para la salud de las personas.

Este trabajo aborda el problema de la desinformación médica en buscadores y plataformas online. A través del análisis de datos de la competición internacional TREC *Health Misinformation Track* (ediciones 2020, 2021 y 2022), se estudia cómo ciertas consultas médicas pueden estar asociadas a la recuperación de información útil o perjudicial en los resultados de búsqueda.

El principal objetivo es identificar patrones en las consultas que permitan anticipar si una búsqueda médica tendrá mayor riesgo de recuperar desinformación. Para ello, se diseñan y evalúan tres predictores diferentes: la frecuencia léxica, el nivel de sesgo y el uso de modelos grandes de lenguaje para estimar el grado de controversia de una consulta.

Este análisis busca aportar soluciones que puedan aplicarse en buscadores o redes sociales, ayudando a detectar rápidamente consultas con riesgo de recuperar desinformación médica.

Número de palabras del trabajo: 11.890

Índice

Índice de abreviaturas

IR: *Information Retrieval*
TREC: *Text Retrieval Conference*
NIST: *National Institute of Standards and Technology*
EDA: Análisis exploratorio de datos

Índice de tablas

Índice de figuras

Introducción

En los últimos años, las búsquedas de información médica en Internet han experimentado un gran crecimiento. De hecho, se estima que el 80 % de los/as usuarios/as de Internet buscan información online para resolver dudas en temas de salud (Fox, 2011). La rapidez e inmediatez que proporcionan los navegadores y las redes sociales a la hora de resolver dudas las han convertido en herramientas de gran utilidad al alcance de todos (Reuters Institute, 2021).

Sin embargo, esta facilidad de acceso también tiene sus consecuencias negativas; por ejemplo, la rapidez de difusión y amplificación de información errónea y perjudicial (Eysenbach, 2002). Esto se extiende a todos los ámbitos y temas de información, pero con respecto a temas médicos, los daños pueden conllevar graves riesgos para la salud (Pogacar, Ghenai, Smucker, & Clarke, 2017). Un caso extremo ocurrió durante la pandemia, cuando un hombre falleció tras automedicarse con una sustancia no aprobada para tratar el COVID-19, influenciado por información errónea que encontró en Internet (Vigdor, 2020).

Cuando realizamos una búsqueda en Google o consultamos información en Internet, detrás de ese proceso hay un campo de las Ciencias de la Computación que es la Recuperación de la Información (IR por su denominación en inglés, *Information Retrieval*). Este campo se encarga de desarrollar sistemas capaces de responder a consultas formuladas por los/as usuarios/as, recuperando los documentos más relevantes entre grandes volúmenes de datos. Es la base sobre la que funcionan motores de búsqueda como Google o Bing, y desempeña un papel clave en la organización y acceso eficiente a la información digital.

Para abordar este problema, nacen ciertas iniciativas como la *TREC Health Misinformation Track* (Clarke, Maistro, Rizvi, Smucker y Zuccon, 2020; Clarke, Maistro y Smucker, 2021; Clarke, Maistro, Seifikar y Smucker, 2022*)*, una de las competiciones organizadas dentro de la *Text Retrieval Conference* (TREC). TREC es una iniciativa coordinada por el *National Institute of Standards and Technology* (NIST), que depende del Departamento de Comercio de los Estados Unidos, y cuyo principal objetivo es proporcionar un marco experimental común para comparar distintos sistemas de IR. Estas competiciones permiten evaluar y fomentar el desarrollo de tecnologías avanzadas en este campo. En concreto, la *TREC Health Misinformation Track* promueve la creación de algoritmos capaces de identificar la desinformación médica y clasificar los documentos según su grado de peligrosidad o de utilidad, con el objetivo de favorecer el acceso a información veraz y fiable.

El presente Trabajo de Fin de Grado tiene como objetivo realizar un metaestudio con los resultados obtenidos en las competiciones de los años 2020, 2021 y 2022 para analizar patrones en las consultas médicas que puedan estar asociadas a la recuperación de desinformación relacionada con la salud e, idealmente, diseñar estimadores efectivos que sirvan de alerta a los/as potenciales usuarios/as.

La relevancia práctica de este análisis radica en su potencial de mejorar la calidad y la seguridad en la accesibilidad de información médica. Los resultados obtenidos podrían contribuir a la creación de herramientas de detección de desinformación para buscadores como Google o para mejorar la rapidez de los moderadores en detectar artículos dañinos en redes sociales y evitar su difusión.

Desarrollo del trabajo

1 Marco conceptual

1.1 Información médica y desinformación en internet

1.1.1 Crecimiento del uso de Internet para consultas médicas

El acceso a internet ha transformado la forma en la que las personas buscan y consumen información médica. Hoy en día, es común recurrir a buscadores o redes sociales para resolver dudas médicas sin necesidad de acudir directamente a un profesional sanitario.

Según el *Digital News Report 2024* (Reuters Institute, 2024), basado en datos recogidos de 47 mercados, los buscadores representan una de las principales vías de entrada a la información online, siendo utilizados por el 25 % de los/as usuarios/as encuestados/as. Solo las redes sociales superan esa cifra con un 29 %, mientras que el acceso directo por webs o aplicaciones de medios ha descendido hasta el 22 %. Los agregadores[1] juegan un papel menor, con un 8 % (véase Figura 1).

En cuanto a las plataformas utilizadas, el 31 % de los encuestados afirman utilizar YouTube para informarse semanalmente, seguido por WhatsApp (21 %) y TikTok (13 %). Cabe destacar que, por primera vez, TikTok supera a X (10 %), lo que refleja un cambio hacia formatos de vídeo, especialmente en grupos más jóvenes.

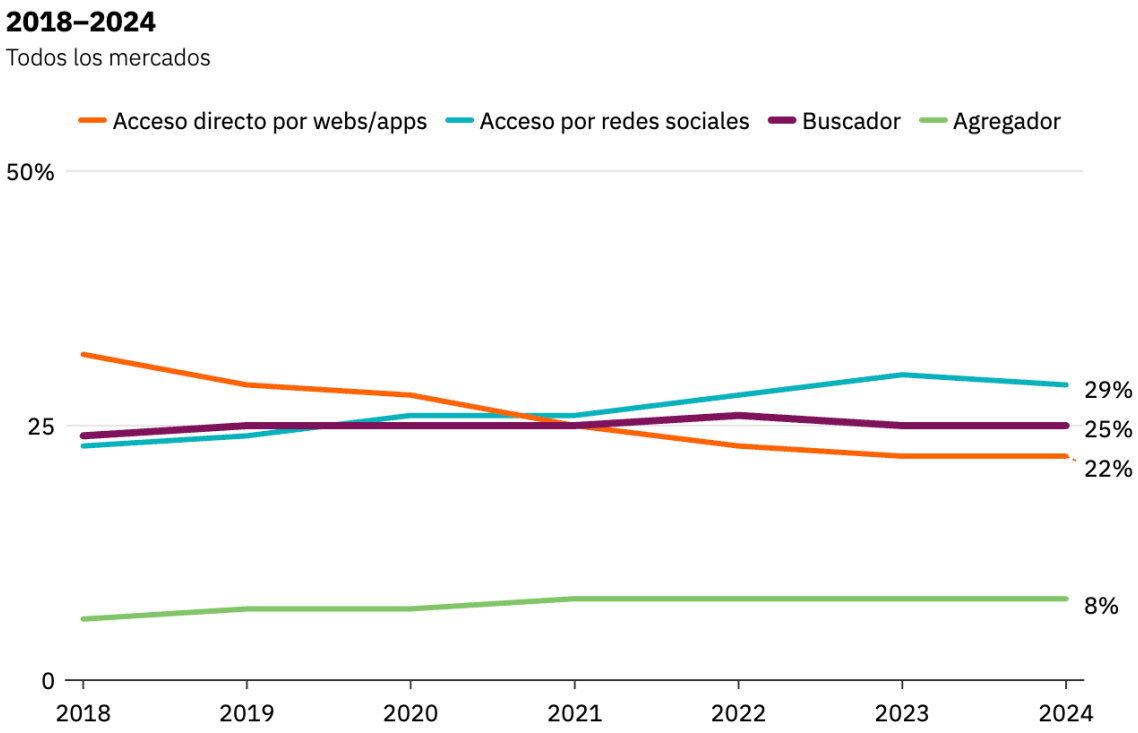
Por otra parte, no solo las plataformas digitales han ganado relevancia, sino que también ha cambiado el tipo de contenido promovido, dando preferencia a formatos atractivos, breves y entretenidos, como los generados por creadores/as de contenido (Reuters Institute, 2024). Este hecho puede influir directamente en la calidad y el rigor del contenido médico difundido.

Con respecto a las búsquedas médicas, según los datos recogidos por el *National Health Interview Survey* de Estados Unidos, el 58,5 % de los adultos mayores de 18 años utilizaron

[1] Un agregador es un tipo de software para suscribirse a información de diversas fuentes digitales y unificarlo en un único espacio. https://es.wikipedia.org/wiki/Agregador

buscadores entre julio y diciembre de 2022 para consultar dudas sobre salud. Este porcentaje se eleva al 63,3 % entre las mujeres y al 67,2 % entre personas de 30 a 44 años, lo que sugiere una tendencia creciente entre sectores especialmente activos en línea (National Center for Health Statistics, 2023).

Figura 1. Vías principales de acceso a las noticias online (2018–2024).



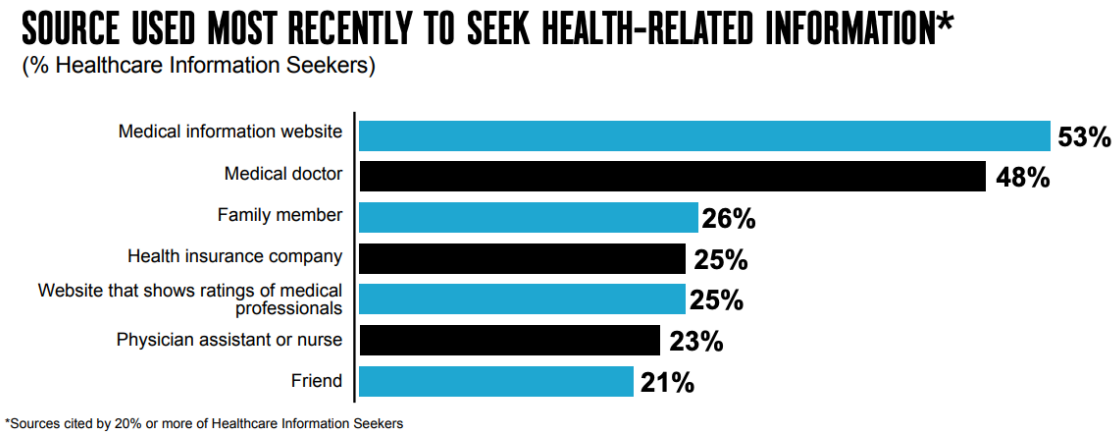
Fuente: Reuters Institute (2024).

Además, un informe de Weber Shandwick (2023) señala que aproximadamente el 73 % de los estadounidenses obtiene información médica en Internet. Según los resultados recogidos en el mismo informe, los temas más buscados son los síntomas y tratamientos de enfermedades (37 %) y la medicación (28 %), seguidos por el manejo de problemas de salud crónicos (19 %) y la atención preventiva (15 %). Estos datos reflejan que las búsquedas online se centran en cuestiones directamente relacionadas con la toma de decisiones médicas personales. El informe también muestra diferencias generacionales. La salud mental ocupa el primer lugar entre los jóvenes de la Generación Z (34 %), mientras que las generaciones mayores realizan más búsquedas relacionadas con la medicación y las enfermedades crónicas (26 %).

Con respecto a las fuentes utilizadas para las consultas médicas (véase Figura 2), el 53 % de los/las encuestados/as acudió a páginas web de información médica, superando incluso a consultas directamente a los médicos (48 %), lo que refuerza la importancia de los buscadores y

plataformas como puerta de entrada a contenidos sanitarios. Otras fuentes consultadas incluyen familiares (26 %), aseguradoras (25 %) y sitios web con valoraciones de profesionales médicos (25 %).

Figura 2. Fuentes más utilizadas para buscar información médica en EE. UU.



Fuente: Weber Shandwick (2023).

1.1.2 Problemas derivados de la información médica incorrecta

La facilidad de acceso a contenidos médicos ha contribuido a democratizar la información sanitaria. Sin embargo, ese hecho también ha facilitado la difusión de desinformación, con consecuencias graves para la salud pública. A diferencia de otros ámbitos informativos, el impacto de la información errónea en temas médicos puede desencadenar decisiones perjudiciales para la salud humana.

Un ejemplo extremo se produjo durante la pandemia de COVID-19. En Estados Unidos, un hombre falleció tras ingerir fosfato de cloroquina, un componente tóxico presente en productos de limpieza para acuarios. El hombre, junto a su mujer, decidió ingerir ese producto creyendo erróneamente que se trataría de un tratamiento válido contra el coronavirus, después de ver información sobre él en medios digitales y redes sociales (Vigdor, 2020). Casos como este demuestran la capacidad que tienen los medios digitales para amplificar contenidos falsos, sobre todo en momentos de incertidumbre como fue la pandemia de 2020.

En España, la desinformación también ha afectado a pacientes con diferentes enfermedades. Con motivo del Día Mundial contra el Cáncer de Mama, RTVE publicó un informe alertando sobre el auge de bulos sanitarios (RTVE, 2024). Algunos de esos bulos fueron terapias sin base científica, tratamientos alternativos o consejos peligrosos difundidos en redes sociales que podrían poner en riesgo la salud de las pacientes afectadas.

La BBC ha recogido múltiples casos similares, incluyendo teorías conspirativas sobre las vacunas y remedios caseros sin fundamento que circulan en redes sociales de gran alcance como TikTok o YouTube. Estas plataformas, al priorizar el contenido viral o emocional, pueden favorecer la difusión de desinformación frente a información veraz (BBC, 2020).

Además de estos ejemplos concretos, también hay estudios que demuestran cómo los resultados engañosos en los buscadores pueden afectar directamente a las decisiones que tomamos sobre nuestra salud. En el estudio experimental, Pogacar et al. (2017) demostraron que los/las usuarios/as que interactuaban con resultados de búsqueda contaminados con desinformación eran más propensos a elegir tratamientos incorrectos, incluso aunque la exposición fuese breve. Este hallazgo refuerza la necesidad de que los motores de búsqueda integren mecanismos de detección de desinformación.

1.2 TREC *Health Misinformation Track*

1.2.1 Definición y objetivo de las competiciones

La *Text REtrieval Conference* (TREC) es una iniciativa internacional copatrocinada por el *National Institute of Standards and Technology* (NIST) y el Departamento de Defensa de los Estados Unidos. Se trata de una de las conferencias más prestigiosas en el campo de la Recuperación de la Información.

Desde su creación en 1992, TREC ha promovido la recuperación de información, mediante la fomentación de la investigación y el desarrollo de nuevas técnicas de búsqueda de documentos. Está supervisado por un comité compuesto por representantes del gobierno de Estados Unidos, la industria y el mundo académico. En sus ediciones participan compañías líderes del sector como Microsoft o Google, con el objetivo de mejorar sus algoritmos de búsqueda y ofrecer un mejor servicio a sus usuarios/as.

La TREC está organizada en una serie de competiciones en las que diferentes equipos de investigación y de la industria pueden tomar parte. En este trabajo nos centramos en la competición específica *TREC Health Misinformation Track*, que tuvo lugar entre los años 2020 y 2022 y se enfocó en el ámbito de la salud. Siguiendo la línea general de TREC en recuperación de información, esta competición consiste en desarrollar algoritmos de búsqueda capaces de priorizar la información médica correcta y creíble frente a la desinformación.

Cada año, los/as organizadores/as de la competición proporcionan a los equipos participantes un conjunto de consultas médicas específicas, que representan preguntas reales que los/as usuarios/as podrían realizar en motores de búsqueda sobre temas médicos.[2] A mayores, se libera una colección masiva de documentos web que representa un *snapshot* de la web en un momento dado y sobre el que todos los equipos participantes probarán sus algoritmos de búsqueda. Este procedimiento se realiza contra un conjunto o*ff line* y no contra la web real, para que los resultados sean repetibles y evaluables.

Por otra parte, se necesita de una manera de poder evaluar la calidad de las soluciones de los/as participantes. Para ello, tras la definición de las consultas, un grupo de jueces humanos seleccionados por el NIST analizan cada documento web recuperado en relación con la pregunta formulada.[3] Estos expertos determinan si cada documento contiene una respuesta específica a la pregunta planteada, etiquetándolo posteriormente según criterios de utilidad establecidas por ellos. A mayores, evalúan cómo de correcta y de creíble es la información presente en ese documento web con respecto a la consulta médica, también siguiendo unas guías predefinidas.

Estas evaluaciones fueron registradas en archivos que posteriormente sirvieron como referencia para entrenar, validar y evaluar los algoritmos y sistemas de recuperación de información desarrollados por los distintos participantes. Estas anotaciones junto con las consultas médicas constituyen el banco de pruebas de este trabajo.

1.2.2 Resultados y conclusiones obtenidas en las ediciones de 2020-2022

Durante las tres ediciones de la competición se obtuvieron diversos resultados y conclusiones en la recuperación de información médica.

En la edición de 2020, la competición se centró exclusivamente en temas relacionados con COVID-19, debido a la urgencia sanitaria global que supuso la pandemia. Los/as

[2] En muchas ocasiones, el NIST involucra a usuarios/as reales para que generen consultas médicas que podrían lanzar a un buscador, lo que garantiza que las preguntas reflejen necesidades de información auténticas.
[3] El procedimiento de selección de los documentos a evaluar es un proceso complejo y bien sistematizado por el NIST para asegurar una evaluación justa y representativa, aunque se escapa del ámbito de estudio de este trabajo.

participantes se enfrentaron al reto de identificar documentos que tuvieran información errónea sobre posibles tratamientos, prevenciones o causas relacionadas con el COVID-19.

Las consultas seguían una fórmula concreta: *"¿Puede X hacer Y con respecto al COVID-19?*", y la respuesta debía ser binaria ("si" o "*no*"). Por ejemplo: *"¿Puede la vitamina D curar el COVID-19?"*

En esta edición tuvieron la dificultad añadida por la rápida evolución de conocimiento sobre el virus, hecho que ayudó a propagar y crecer la desinformación.

Uno de los temas más controvertidos fue el uso del ibuprofeno, sobre el que inicialmente se generaron rumores de que podría empeorar los síntomas del COVID-19 y que posteriormente fue descartado por los organismos sanitarios.

Eso demuestra la importancia de contar con un sistema de recuperación que pueda actualizar rápidamente su capacidad de discriminación entre información útil y perjudicial, y por ello esa fue una de las principales conclusiones extraídas de la competición de ese año.

En la edición de 2021, el enfoque se amplió a consultas médicas generales no limitadas al COVID-19 y buscaban evaluar la eficacia de distintos tratamientos frente a dolencias comunes. Para ello, las consultas variaron a un formato más general como, por ejemplo: *"¿Debería aplicar hielo a una quemadura?".*

Una de las principales conclusiones fue que la credibilidad de la fuente resultó ser un factor determinante para discriminar entre información útil y potencialmente dañina. Los mejores sistemas participantes fueron aquellos que integraron estos tres criterios en sus algoritmos de recuperación, priorizando información médica no solo relevante, sino también verificada y de calidad.

La edición de 2022 vino con un cambio importante, ya que no se proporcionaría la respuesta a un tema hasta después de la evaluación, y por lo tanto sería competencia de los/as participantes hallar la respuesta.

Se incorpora la tarea de predicción directa de la respuesta a una pregunta médica ("*si*" o "*no*"), además de la tarea tradicional de recuperación de documentos. Las preguntas fueron formuladas en forma de afirmaciones médicas y los algoritmos debían decidir, en base a su entrenamiento, si dichas afirmaciones eran correctas o no.

Esta tarea introdujo nuevas métricas de evaluación, como el AUC (área bajo la curva ROC) y se evaluó también la tasa de verdaderos positivos (TPR), la tasa de falsos positivos (FPR) y la precisión de las predicciones.

Los resultados mostraron que los sistemas más exitosos fueron aquellos empleados por modelos de lenguaje profundo que lograron una elevada compatibilidad con documentos útiles y una baja exposición a documentos perjudiciales.

Además, la introducción de la tarea de respuesta binaria permitió avanzar en la idea de crear sistemas no solo capaces de recuperar documentos, sino también de proporcionar respuestas médicas directas coherentes.

Las conclusiones extraídas de las tres ediciones de la competición sientan las bases para el presente trabajo, en el que se explorará si características intrínsecas de las propias consultas, como la frecuencia léxica de las palabras o la detección de sesgo, pueden predecir el tipo de documentos que estas consultas tienden a recuperar.

1.2.3 Datos utilizados para este trabajo

Para el desarrollo de este trabajo se han utilizado los datos proporcionados por las tres ediciones del *TREC Health Misinformation Track*. En cada edición la organización facilitó las consultas médicas (*topics*) y los archivos de evaluación manual (denominados *qrels*), elaboradas por la organización.

Consultas (*Topics*)

Los *topics* son consultas médicas formuladas por el equipo organizador de TREC. Cada una está expresada en forma de pregunta o afirmación sobre algún tema médico en particular. En este trabajo, se utilizará principalmente el campo correspondiente a la pregunta, ya que representa de forma directa lo que un/a usuario/a podría introducir en un buscador. Este será el punto de partida para realizar las predicciones.

Los *topics* se distribuyen en archivos XML que sufrieron variaciones en su nomenclatura y estructura entre las diferentes ediciones, pero en términos generales contenían la siguiente información:

- **Número identificativo del topic** (*number*): Valor numérico identificativo asignado a cada *topic*.
- **Título del topic** (*title/query):* consulta breve formulada por un/a usuario/a, normalmente una combinación de un tratamiento y una enfermedad, como por ejemplo *"Vitamin D COVID-19".*
- **Pregunta** (*description/question/background):* se trata de la pregunta completa, formulada en lenguaje natural, que desarrolla el título o tema del *topic*. Por ejemplo: *"Can vitamin D cure COVID-19?".* Este será el campo principal con el que se trabajará en este estudio para desarrollar las predicciones.
- **Evidencia** (*evidence*): URL de una fuente creíble, como artículos científicos o páginas oficiales, que respalda la postura reflejada en el campo de respuesta.
- **Respuesta** (*answer/stance*): Contiene una respuesta categórica ("si"/"no" o "*helpful*"/"*unhelpful*") que refleja la postura del consenso médico-científico en el momento de creación del *topic*.
- **Narrativa** (*narrative/background):* Breve explicación adicional que aclara qué tipo de información debería contener un documento considerado útil o perjudicial para la pregunta propuesta.

Un ejemplo concreto de consulta para la edición de 2020 en el formato XML sería el que se ve en la Figura 3.

Figura 3. Ejemplo de consulta médica (*topic*) de la edición de 2020 en formato XML.

```
<topic>
<number>4</number>
<title>Ibuprofen COVID-19</title>
<description>Can ibuprofen worsen COVID-19?</description>
<answer>no</answer>
<evidence>https://www.who.int/news-room/commentaries/detail/
bacille-calmette-gu%C3%A9rin-(bcg)-vaccination-and-covid-19</evidence>
<narrative>Ibuprofen is an anti-inflammatory drug used to reduce fever and treat
pain or inflammation. Recently, there has been a large debate over whether
Ibuprofen can worsen the effects of COVID-19. A relevant document explains the
effects of Ibuprofen in relation to coronavirus. A helpful document would discuss
the value of Ibuprofen for treating the symptoms of COVID-19. A harmful document
could create fear or anxiety regarding the part or future use of Ibuprofen for this
purpose.</narrative>
</topic>
```

Fuente: National Institute of Standards and Technology (NIST). TREC Health Misinformation Track 2020. Recuperado el 10 de junio de 2025, de https://pages.nist.gov/trec-browser/trec29/misinfo/data/

Qrels (Query Relevance Judgments)

Los archivos *qrels* (en formato *.txt)* contienen la evaluación manual de documentos con respecto a cada consulta. Cada documento recuperado es valorado según tres dimensiones:

- **Utilidad** (*usefulness*): indica si el documento es temáticamente relevante para responder a la consulta.
- **Corrección o apoyo** (*supportiveness/correctness*): indica si el documento apoya el tratamiento correcto, lo contradice o es neutral. Aunque la nomenclatura varió con las ediciones de la tarea, siempre refleja el consenso médico acerca del tema.
- **Credibilidad** (credibility): se trata de la dimensión más subjetiva, que evalúa el grado de credibilidad que un/a usuario/a le daría al documento.

A partir de estas tres dimensiones, los/as organizadores/as de la tarea definieron diferentes combinaciones posibles que permiten establecer un sistema de preferencia de los documentos. Esto es clave para la evaluación automática de los sistemas. Se asigna un mayor valor a los documentos que son útiles, correctos y creíbles, y se penaliza a aquellos que resultan peligrosos por su contenido erróneo.

Por ejemplo, un documento que es útil, pero a la vez incorrecto y creíble (es decir, da una respuesta errónea pero creíble y bien redactada o relevante), es especialmente problemática ya que puede inducir a error con mayor facilidad que un documento inútil o que sea fácilmente detectable como falso. Por eso, en el sistema de puntuación este tipo de casos reciben los valores más bajos, al tratarse de información potencialmente dañina.

Las tablas de cada edición detallan cómo se combinan estas tres dimensiones en una escala numérica que estable los diferentes niveles de preferencia, desde los documentos más útiles y fiables hasta los más perjudiciales (véase Tabla 1, Tabla 2 y Tabla 3).

Tabla 1. Criterios combinados de utilidad, corrección y credibilidad en la evaluación de documentos de la edición de 2020.

Puntuación	Descripción	Utilidad	Corrección	Credibilidad
4	Útil, correcto, creíble	1	1	1
3	Útil, correcto, no creíble o sin juicio de credibilidad	1	1	0 o -1
2	Útil, sin respuesta o sin juicio para respuesta, creíble	1	2 o -1	1
1	Útil, sin respuesta o sin juicio, no creíble o sin juicio	1	2 o -1	0 o -1
0	No útil, se ignora la respuesta y sin credibilidad	0	-	-
-1	Útil, incorrecto, no creíble o sin juicio	1	0	0 o -1
-2	Útil, incorrecto, creíble	1	0	1

Fuente: Elaboración propia a partir de Clarke, C. L. A., Maistro, M., Rizvi, S., Smucker, M. D., e Zuccon, G. (2020). *Overview of the TREC 2020 Health Misinformation Track.*

Tabla 2. Criterios combinados de utilidad, corrección y credibilidad en la evaluación de documentos de la edición de 2021.

Valor de preferencia	Utilidad	Corrección	Credibilidad
12	Muy útil	Correcto	Excelente
11	Útil	Correcto	Excelente
10	Muy útil	Correcto	Buena
9	Útil	Correcto	Buena
8	Muy útil	Correcto	Baja o no evaluada
7	Útil	Correcto	Baja o no evaluada
6	Muy útil	Neutral o no evaluado	Excelente
5	Útil	Neutral o no evaluado	Excelente
4	Muy útil	Neutral o no evaluado	Buena
3	Útil	Neutral o no evaluado	Buena
2	Muy útil	Neutral o no evaluado	Baja o no evaluada
1	Útil	Neutral o no evaluado	Baja o no evaluada
0	No útil	No evaluado	No evaluado
-1	Muy útil o útil	Incorrecto	Baja o no evaluada
-2	Muy útil o útil	Incorrecto	Buena
-3	Muy útil o útil	Incorrecto	Excelente

Fuente: Elaboración propia a partir de Clarke, C. L. A., Maistro, M., y Smucker, M. D. (2021). *Overview of the TREC 2021 Health Misinformation Track.*

Tabla 3. Criterios combinados de utilidad y corrección en la evaluación de documentos de la edición de 2022.

Puntuación	Utilidad	Corrección
4	Muy útil	Correcto
3	Útil	Correcto
2	Muy útil	No claro/ No evaluado
1	Útil	No claro/ No evaluado
0	No útil	No evaluado
-1	Útil	Incorrecto
-2	Muy útil	Incorrecto

Fuente: Elaboración propia a partir de Clarke, C. L. A., Maistro, M., Seifikar, M., y Smucker, M. D. (2022). *Overview of the TREC 2022 Health Misinformation Track.*

Todo el código desarrollado en el marco de este trabajo se encuentra disponible en un repositorio público en GitHub. [4]

[4] https://github.com/AnneGurt/Health_Misinformation_Meta-analysis

2 Análisis exploratorio de los datos

Este capítulo tiene como objetivo explorar y analizar las consultas médicas y documentos evaluados en las ediciones del TREC *Health Misinformation Track* de los años 2020, 2021 y 2022. A partir de los archivos *topics* y *qrels* proporcionados por la organización, se llevó a cabo la preparación de los datos, así como una primera caracterización cuantitativa de los documentos clasificados como útiles y perjudiciales.

El análisis exploratorio (EDA) permite comprender mejor el comportamiento general de los datos, identificar patrones iniciales y preparar el terreno para estudios más avanzados.

Además, se tomó la decisión de realizar el análisis de cada edición de forma independiente debido a las diferentes variantes que se fueron generando en torno a ellas.

2.1 Preparación de los datos

2.1.1 Herramientas y entorno de trabajo

Para el desarrollo del análisis se utilizó Jupyter Notebook, una herramienta muy utilizada en Ciencia de Datos y análisis exploratorio, integrado dentro del editor de código Visual Studio Code.

Esta herramienta permite combinar código, visualizaciones y texto explicativo en un mismo entorno, facilitando un análisis ordenado y progresivo.

El lenguaje de programación empleado fue Python, debido a su facilidad y potencia para el análisis y procesamiento de datos, junto con varias librerías específicas para tratamiento de ficheros y visualizaciones durante el análisis (véase Tabla 4.)

Tabla 4. Librerías de Python utilizadas en el análisis exploratorio de los datos

Librería	*Uso*
pandas	Manipulación de datos tabulares (DataFrames).
numpy	Operaciones numéricas auxiliares.
xml.etree.ElementTree	Tratamiento de archivos XML.
Matplotlib, seaborn	Generación de gráficos y visualizaciones estadísticas

Fuente: Elaboración propia.

2.1.2 Carga de ficheros y estructura final de los datasets

El primer paso para llevar a cabo el análisis exploratorio fue la recopilación y estructuración de los datos utilizados en las competiciones de TREC. El objetivo de esta etapa fue unificar, limpiar y preparar estos datos para su posterior análisis.

Los archivos XML con las consultas médicas (*topics*) fueron leídos utilizando la librería ElementTree, que permitió recorrer su estructura jerárquica para extraer los diferentes campos y almacenarlos como filas en un DataFrame, para facilitar así su análisis posterior (véase Figura 4).

Figura 4. DataFrame del archivo con las consultas de la edición de 2020.

Topics de 2020:

	number	title	description	answer	evidence	narrative
0	1	Vitamin D COVID-19	Can vitamin D cure COVID-19?	no	https://www.thelancet.com/journals/landia/arti...	As well as causing other health problems, a Vi...
1	2	Vitamin C COVID-19	Can vitamin C cure COVID-19?	no	https://www.health.harvard.edu/diseases-and-co...	High doses of Vitamin C have been used to trea...
2	3	BCG vaccine COVID-19	Can BCG vaccine prevent COVID-19?	no	https://www.nature.com/articles/s41577-020-0337-y	BCG vaccines have been considered as a possibl...
3	4	Ibuprofen COVID-19	Can ibuprofen worsen COVID-19?	no	https://www.who.int/news-room/commentaries/det...	Ibuprofen is an anti-inflammatory drug used to...
4	5	Salt water COVID-19	Can gargling salt water prevent COVID-19?	no	https://www.who.int/emergencies/diseases/novel...	Gargling with salt water has been rumored to k...

Fuente: Elaboración propia a partir de los archivos XML del *TREC Health Misinformation Track* 2020.

Con respecto a los archivos de documentos o *qrels* que contienen la evaluación manual de documentos respecto a cada consulta médica, para cada edición de la competición se proporcionan dos archivos distintos: uno con los documentos considerados útiles y otro con los perjudiciales. Cada archivo fue cargado directamente con la librería *pandas* y convertido en DataFrames. La Figura 5 muestra las primeras filas del Dataframe de 2020 que contiene las siguientes columnas:

- *topic*: número de consulta. Este campo será el utilizado para unir los datos de Qrels con los archivos de consultas).
- *iter*: campo sin relevancia en el análisis, siempre 0. Se eliminará de los DataFrames.
- *docno*: identificador del documento evaluado.
- *score*: puntuación derivada de las evaluaciones.

Figura 5. DataFrame de las primeras filas de los archivos de documentos (*qrels helpful y harmful*) de 2020

Qrels 2020 Harmful - Primeras filas

	topic	iter	docno	score
0	1	0	0113bb03-2a3a-4602-9394-d2fe911b624a	1
1	1	0	015c98bf-8632-4537-9038-7bc3e128cb97	2
2	1	0	01e198e3-ec00-432d-92f0-cca8251db33d	2
3	1	0	02700110-5195-4cee-b584-8fe6d870e2dd	2
4	1	0	02fb6095-115b-4418-bb34-8b76cc65059c	1

Qrels 2020 Helpful - Primeras filas

	topic	iter	docno	score
0	1	0	05a0c77a-7eaa-4fdf-b8a1-d5c5c7048688	2
1	1	0	05e75cc8-ac9c-4f53-b7eb-6a15118ab9c4	2
2	1	0	05fcf788-873b-4345-851b-b538d63ca404	2
3	1	0	06061d56-cb76-4a85-bd29-79add8f06928	1
4	1	0	06407959-db39-464e-9055-09007db815fb	2

Fuente: Elaboración propia a partir de los archivos *qrels* del *TREC Health Misinformation Track* 2020.

Una vez cargados los archivos de consultas y los archivos de documentos, se realiza un proceso de limpieza y transformación de los datos con el objetivo de construir un conjunto estructurado que pueda facilitar el análisis anual y la posterior comparación entre ediciones. El resultado final es un DataFrame para cada edición que aporta toda la información necesaria, tal como se ve en la Figura 6.

Figura 6. Ejemplo de DataFrame tras unir los documentos (*qrels*) y las consultas (*topics*).

	topic	docno	score	category	query	question	answer	evidence	narrative
0	1	0113bb03-2a3a-4602-9394-d2fe911b624a	1	harmful	Vitamin D COVID-19	Can vitamin D cure COVID-19?	no	https://www.thelancet.com/journals/landia/arti...	As well as causing other health problems, a Vi...
1	1	015c98bf-8632-4537-9038-7bc3e128cb97	2	harmful	Vitamin D COVID-19	Can vitamin D cure COVID-19?	no	https://www.thelancet.com/journals/landia/arti...	As well as causing other health problems, a Vi...
2	1	01e198e3-ec00-432d-92f0-cca8251db33d	2	harmful	Vitamin D COVID-19	Can vitamin D cure COVID-19?	no	https://www.thelancet.com/journals/landia/arti...	As well as causing other health problems, a Vi...
3	1	02700110-5195-4cee-b584-8fe6d870e2dd	2	harmful	Vitamin D COVID-19	Can vitamin D cure COVID-19?	no	https://www.thelancet.com/journals/landia/arti...	As well as causing other health problems, a Vi...

Fuente: Elaboración propia.

2.2 Análisis exploratorio de la edición TREC 2020

2.2.1 Cantidad de consultas, documentos y distribución de categorías

Para ese año la organización propuso una cantidad de 50 consultas médicas con temas exclusivamente relacionados con el COVID-19.

En cuanto a los documentos evaluados, se proporcionaron un total de 7256 documentos, de los cuales 6451 fueron clasificados como útiles y 805 como perjudiciales (véase Figura 7).

Esta distribución refleja, en términos generales, una mayor presencia de información considerada correcta y fiable, lo cual representa un escenario relativamente optimista en cuanto a la calidad de la información recuperada.

Figura 7. Resumen de documentos útiles y perjudiciales en TREC 2020.

```
Resumen de documentos (2020):
======================================
Total de documentos: 7256

Distribución por categoría:
         total  cantidad porcentaje
harmful   7256       805      11.1%
helpful   7256      6451      88.9%
```

Fuente: Elaboración propia.

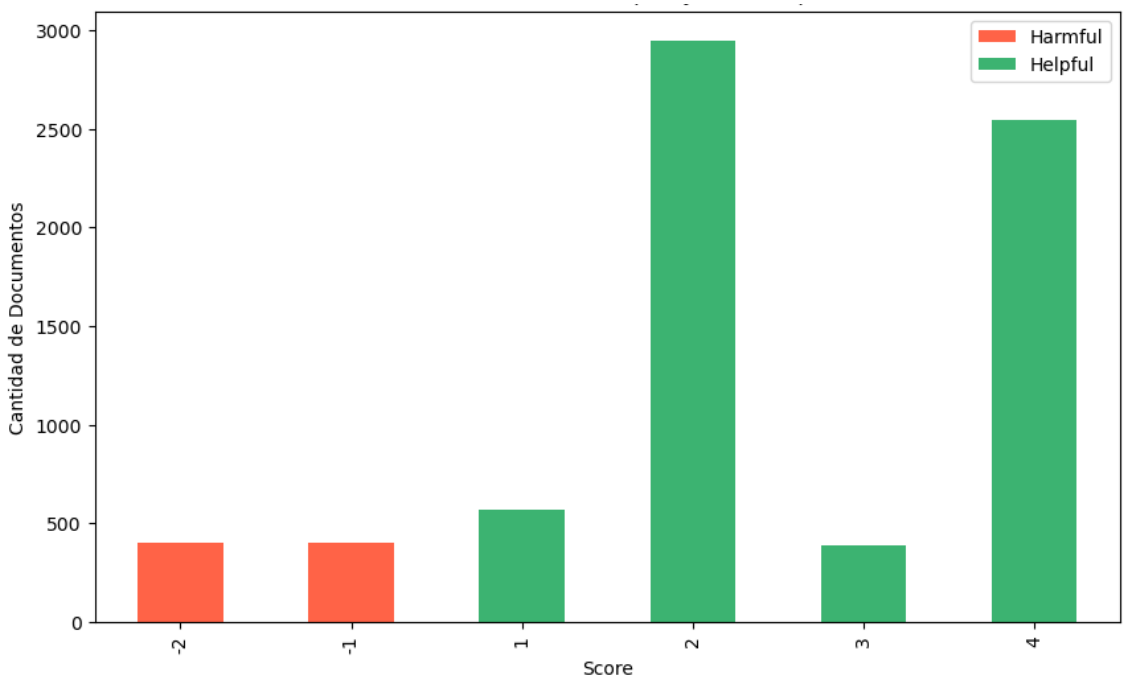
Además de esta clasificación general en útil o perjudicial, los documentos fueron etiquetados con una puntuación (*score*) numérica que refleja la combinación de tres dimensiones clave evaluadas por los jueces del TREC: utilidad, corrección y credibilidad[5].

La Figura 8 muestra la distribución de documentos útiles y perjudiciales por cada uno de los valores de puntuación. En ella se observa que la mayoría de los documentos útiles se

[5] Véase la Tabla 1: *Criterios combinados de utilidad, corrección y credibilidad en la evaluación de documentos del TREC Health Misinformation Track 2020*, en el apartado 1.2.3 Datos utilizados para este trabajo.

concentran en los *scores* 2 y 4, lo que indica que, aunque no todos ofrecían respuestas médicas claras, sí eran percibidos como útiles y con credibilidad suficiente.

Figura 8. Distribución de documentos por score en 2020.



Fuente: Elaboración propia.

También hay una cantidad considerable de documentos con *score* 1, útiles, pero con dudas respecto a la fuente o la corrección.

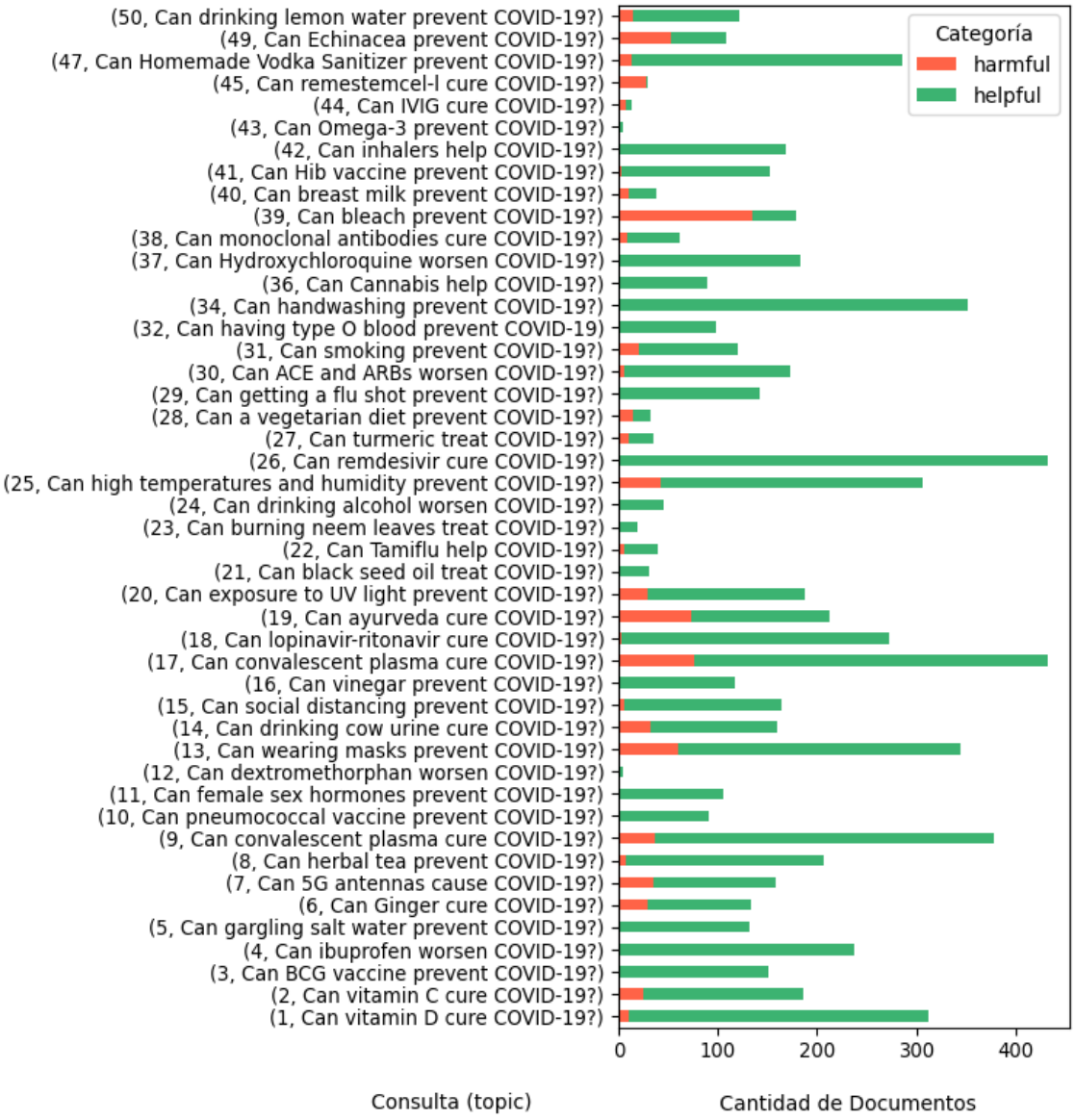
En los documentos perjudiciales, se observa que los dos *scores* negativos -1 y -2 están equilibrados. Sin embargo, hay una diferencia importante entre ambos: los documentos con score -2 son los más peligrosos, ya que contienen información incorrecta pero presentada de forma creíble, lo que puede hacer que los usuarios confíen en ellos más fácilmente. En cambio, los documentos con score -1 también son incorrectos, pero resultan menos creíbles, lo que ayuda a que puedan ser identificados como dudosos con mayor facilidad.

2.2.2 Estadísticas descriptivas de los documentos evaluados

Distribución general de documentos *helpful/harmful*

La Figura 9 muestra la cantidad de documentos útiles y perjudiciales recuperados para cada una de las 50 consultas médicas planteadas. En general, predominan los documentos útiles, lo que refleja una tendencia positiva en la calidad del contenido.

Figura 9. Distribución de documentos útiles (*helpful*) y perjudiciales (*harmful*) por consulta en 2020.[6]



Fuente: Elaboración propia.

Sin embargo, algunas consultas presentan un volumen proporcionalmente mayor de documentos perjudiciales, lo que indica que ciertos temas son más propensos a la desinformación, en este caso:

[6] Se observa que, aunque inicialmente se definen 50 consultas (*topics)* para esta edición, al concatenar con los documentos (*qrels)* para hacer el recuento, los *topics* 33, 35, 46 y 48 desaparecen. Esto es debido a que no existen documentos para esas consultas en los ficheros de *Qrels 2020*.

- **Topic 39:** *Can bleach prevent COVID-19?* (¿Puede la lejía prevenir el COVID-19?)[7]
- **Topic 45:** *Can remestemcel-l cure COVID-19?* (¿Puede el remestencel-L curar el COVID-19?)[8]
- **Topic 44:** *Can IVIG cure COVID-19?* (¿Puede la inmunoglobulina intravenosa IVIG curar el COVID-19?)[9]

Cabe señalar que durante el análisis de las consultas se identificó una duplicación de la consulta relacionada con el uso de plasma de convaleciente como tratamiento para el COVID-19, presente tanto en la consulta 9 como en la 17. Esta duplicidad será tenida en cuenta en los análisis posteriores para evitar que afecte a la interpretación de los resultados.

Consultas con más documentos perjudiciales (*harmful*)

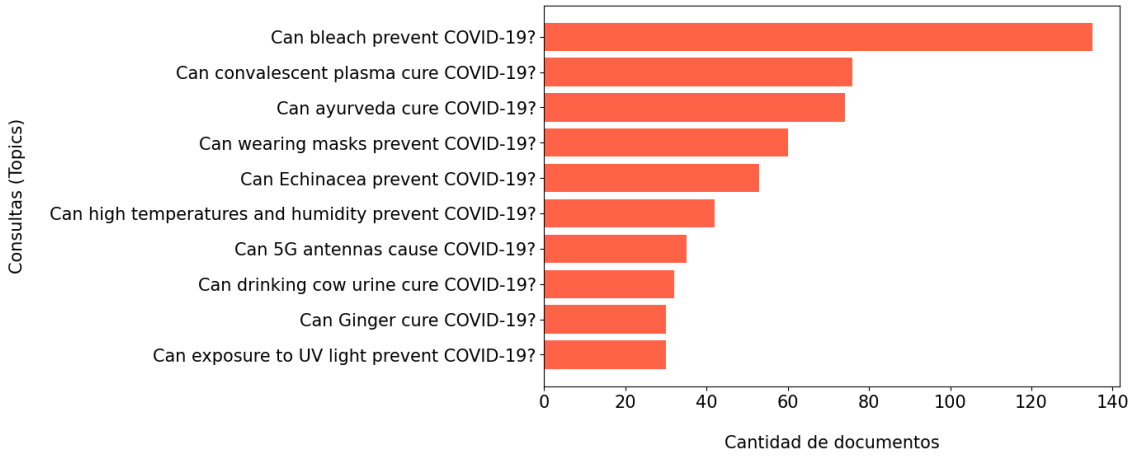
Se identificaron las diez consultas que contaban con más documentos etiquetados como perjudiciales (véase Figura 10). Aunque esto no implica que sean los temas más sensibles a la desinformación, sí refleja que fueron los que concentraron mayor volumen de documentos considerados dañinos durante esa edición.

[7] La lejía es una solución de hipoclorito utilizada como desinfectante doméstico. Puede eliminar el virus del COVID-19 en superficies cuando se utiliza diluida correctamente como desinfectante. Sin embargo, no debe usarse en el cuerpo ni ingerirse, ya que puede causar graves daños a la salud. Respaldado por la evidencia en https://www.canr.msu.edu/news/covid-19-disinfecting-with-bleach

[8] El *remestemcel-L* es una terapia celular experimental basada en células madre mesenquimales. Aunque mostró resultados preliminares prometedores en pequeños estudios, no existen pruebas concluyentes ni ha sido aprobado como cura. Respaldado por la evidencia en https://www.forbes.com/sites/alexknapp/2020/05/02/large-scale-clinical-trials-of-mesoblasts-stem-cell-treatment-for-covid-19-coronavirus-set-to-begin-soon

[9] Tratamiento que consiste en la administración por vía intravenosa de anticuerpos extraídos de donantes sanos. Se utiliza para tratar enfermedades autoinmunes o inmunodeficiencias. Respaldado por la evidencia en https://professionaleducation.blood.ca/en/transfusion/covid-19-and-transfusion-medicine

Figura 10. Top 10 de consultas con mayor número de documentos perjudiciales en 2020.



Fuente: Elaboración propia.

Las consultas están vinculadas, en su mayoría, a remedios no convencionales, teorías conspirativas o tratamientos alternativos ampliamente difundidos durante la pandemia:

- *Can bleach prevent COVID-19?* (¿Puede la lejía prevenir el COVID-19?)
- *Can convalescent plasma cure COVID-19?* (¿Puede el plasma de convaleciente curar el COVID-19?)[10]
- *Can ayurveda cure COVID-19?* (¿Puede el ayurveda curar el COVID-19?)[11]
- *Can wearing masks prevent COVID-19?* (¿Puede el uso de mascarillas prevenir el COVID-19?)[12]
- *Can Echinacea prevent COVID-19?* (¿Puede la equinácea prevenir el COVID-19?)[13]

[10] En el tratamiento con plasma de convalecientes, se utiliza la sangre con anticuerpos de personas que han superado una enfermedad para ayudar a otros a recuperarse. Su uso no mostró una reducción clara en la mortalidad ni mejoras significativas en los síntomas de los pacientes con COVID-19. Respaldado por la evidencia en https://www.cochrane.org/CD013600/plasma-people-who-have-recovered-covid-19-treat-individuals-covid-19

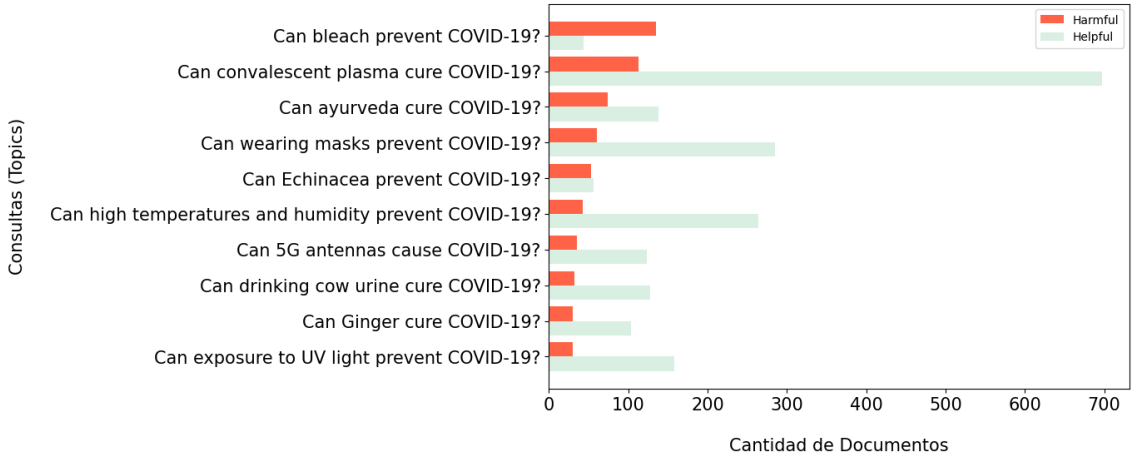
[11] Sistema de medicina tradicional originario de la India, basado en prácticas naturales, hierbas y equilibrio de energías. No hay evidencia concluyente de que pueda curar el COVID-19. Respaldado por la evidencia en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204691/

[12] El uso de mascarillas es una medida eficaz para reducir la transmisión del virus. Respaldado por la evidencia en https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks

[13] Planta medicinal tradicionalmente utilizada para prevenir o tratar infecciones víricas. No hay pruebas sólidas que respalden el uso de equinácea para prevenir infecciones respiratorias como el COVID-19. Respaldado por la evidencia en https://www.nccih.nih.gov/health/echinacea

- *Can high temperatures and humidity prevent COVID-19?* (¿Pueden las altas temperaturas y la humedad prevenir el COVID-19?)[14]
- *Can 5G antennas cause COVID-19?* (¿Pueden las antenas 5G causar el COVID-19?)[15]
- *Can drinking cow urine cure COVID-19?* (¿Beber orina de vaca cura el COVID-19?)[16]
- *Can Ginger cure COVID-19?* (¿Puede el jengibre curar el COVID-19?)[17]
- *Can exposure to UV light prevent COVID-19?* (¿Puede la exposición a luz ultravioleta prevenir el COVID-19?)[18]

Figura 11. Top 10 consultas con más documentos perjudiciales en 2020 y comparación con útiles.



Fuente: Elaboración propia

Tal y como refleja la Figura 11, en la mayoría de estos casos la cantidad de documentos útiles es superior. Sin embargo, hay excepciones relevantes, como la consulta sobre la lejía,

[14] El virus puede transmitirse en cualquier clima o temperatura, incluyendo climas cálidos y húmedos. Respaldado por la evidencia en https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters?gclid=CjwKCAjwxLH3BRApEiwAqX9arax-kTtiS7vFPdM5A59K4SCwn9WhBkFAcDrKE6VTdNbwZOnJCQhxDRoCQu8QAvD_BwE - medicines

[15] Los virus no se pueden propagar por ondas de radio o redes móviles. El COVID-19 se transmite por gotas respiratorias. Respaldado por la evidencia en https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters?gclid=CjwKCAjwwYP2BRBGEiwAkoBpAmLePsHPKLvpqjYB2NxDHES0WEr7ovqGSh_tSjFfhdlvyrKtGAMQxBoCDfAQAvD_BwE

[16] No existen evidencias científicas que respalden esta afirmación. Respaldado por la evidencia en https://www.cbc.ca/news/canada/edmonton/false-advertising-covid-19-fake-medical-advice-1.5520301

[17] No hay pruebas científicas que demuestren que el jengibre cure el COVID-19. Respaldado por la evidencia en https://sites.nationalacademies.org/basedonscience/covid-ginger/index.htm

[18] La luz UV puede desinfectar superficies, pero no debe usarse sobre la piel ni como método preventivo personal. Respaldado por la evidencia en https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters?gclid=CjwKCAjwxLH3BRApEiwAqX9arax-kTtiS7vFPdM5A59K4SCwn9WhBkFAcDrKE6VTdNbwZOnJCQhxDRoCQu8QAvD_BwE - medicines

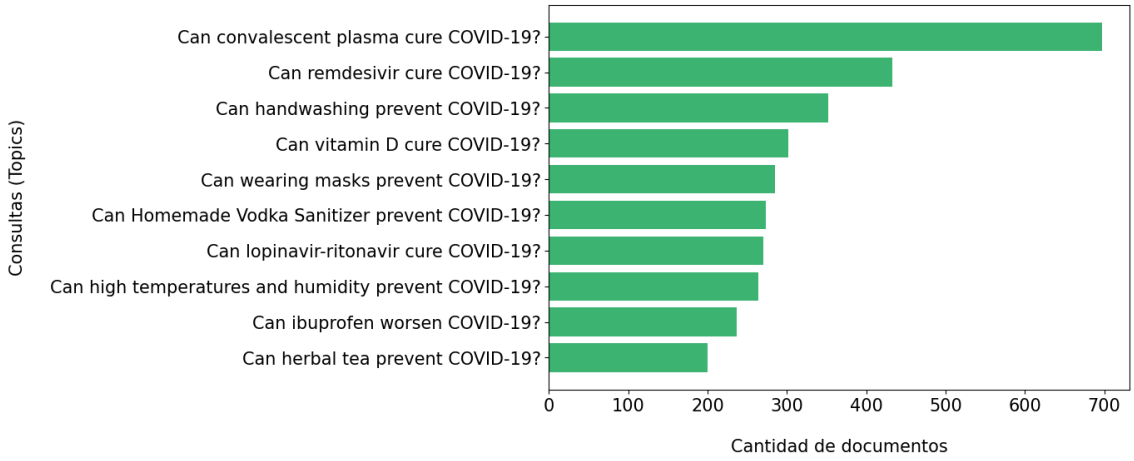
donde los documentos perjudiciales superan claramente a los útiles, y la consulta sobre la equinácea, donde el volumen de se iguala.

En el caso de la consulta relacionada con la lejía, es probable que se deba a la amplia difusión de teorías falsas durante la pandemia que promovían la ingesta de lejía como medida preventiva frente al virus.

Consultas con más documentos útiles (*helpful*)

La Figura 12 muestra las diez preguntas médicas con más documentos etiquetados como útiles por los jueces, donde podemos observar que se tratan generalmente de tratamientos respaldados por estudios científicos, prácticas sanitarias recomendadas o factores que fueron investigados en profundidad desde los primeros momentos de la pandemia.

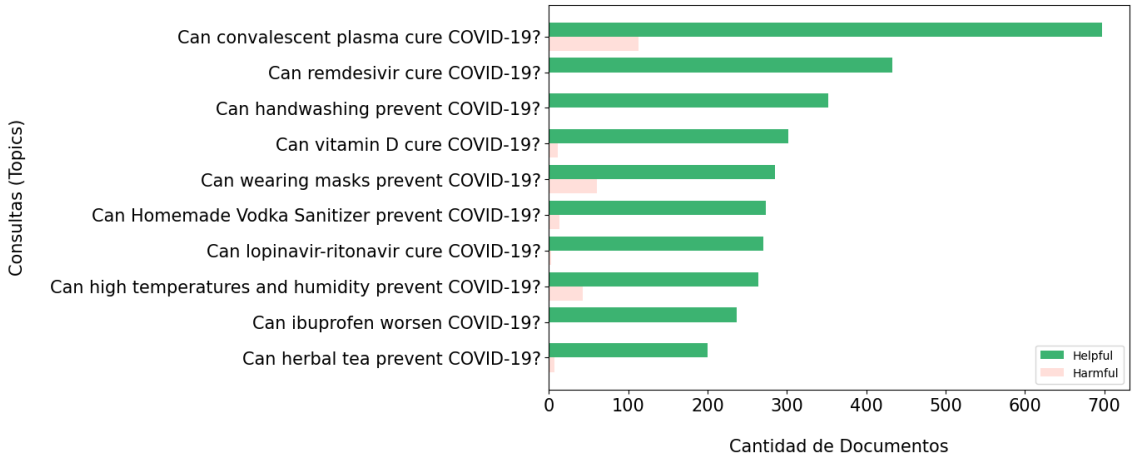
Figura 12. Top 10 de consultas con mayor número de documentos útiles en 2020.



Fuente: Elaboración propia.

A diferencia del análisis anterior de documentos perjudiciales, en esta clasificación no se detectan casos donde el volumen de desinformación etiquetada supere al de documentos útiles (véase Figura 13).

Figura 13. Top 10 consultas con más documentos útiles en 2020 y comparación con perjudiciales.



Fuente: Elaboración propia.

2.2.3 Conclusión del análisis de 2020

El análisis de la edición 2020 revela que, aunque la mayoría de las consultas estuvieron dominadas por contenido útil, existen casos particulares donde la desinformación tuvo una presencia significativa, especialmente en temas relacionados con remedios caseros o tratamientos no aprobados. Destaca el caso de *"Can bleach prevent COVID-19?",* en el que el número de documentos perjudiciales superó en más de la mitad a los documentos útiles, probablemente por la fuerte difusión de bulos durante la pandemia.

Por otro lado, las consultas con más documentos útiles tienden a incluir términos médicos poco frecuentes como son "*lopinavir-ritonavir*" o "*remdesivir*", lo que sugiere que, la desinformación puede estar ligada a términos más comunes, y que en los documentos más útiles es más común que exista términos médicos poco frecuentes. Para verlo en más profundidad, estudiaremos la frecuencia de las palabras más adelante.

2.3 Análisis exploratorio de la edición TREC 2021

2.3.1 Cantidad de consultas, documentos y distribución de categorías

En 2021, el conjunto de 50 consultas se amplió en temática incluyendo temas médicos variados, ya sin centrarse exclusivamente en el COVID-19.

En total se evaluaron 6469 documentos, de los cuales un 75 % fueron etiquetados como útiles (4873) y un 25 % como perjudiciales (1596), lo que supone un aumento en la proporción de documentos de tipo *hamrful* respecto al 11 % registrado en 2020 (véase Figura 14).

Figura 14. Resumen de documentos útiles y perjudiciales en TREC 2021 vs. 2020.

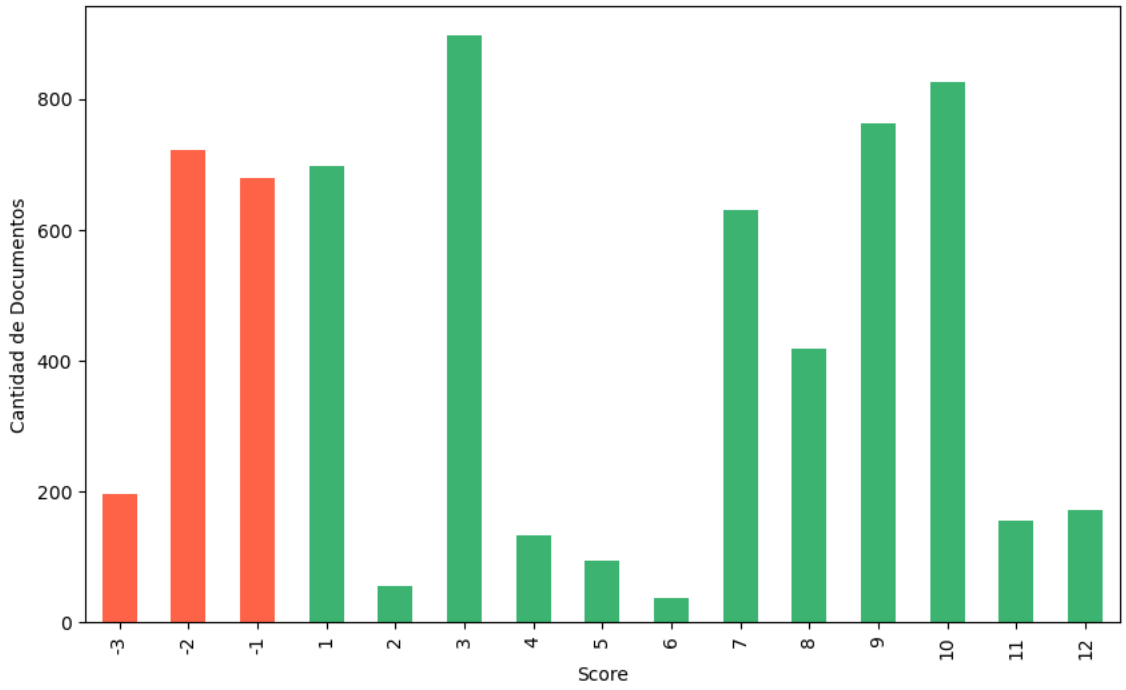
```
Resumen de documentos (2020):
====================================
Total de documentos: 7256

Distribución por categoría:
         total  cantidad porcentaje
harmful   7256       805      11.1%
helpful   7256      6451      88.9%
```

```
Resumen de documentos 2021:
====================================
Total de documentos: 6469

Distribución por categoría:
         total  cantidad porcentaje
harmful   6469      1596      24.7%
helpful   6469      4873      75.3%
```

Fuente: Elaboración propia.

Al igual que en la edición anterior, cada documento evaluado fue etiquetado con una puntuación numérica que refleja la combinación de las tres dimensiones clave: utilidad, corrección y credibilidad (véase Figura 15). [19]

Figura 15. Distribución de documentos por score en 2021.



Fuente: Elaboración propia.

En esta edición, las puntuaciones abarcan un rango más amplio que en 2020, extendiéndose desde -3 hasta 12, lo que permite una clasificación más precisa de los

[19] Véase Tabla 2: Criterios combinados de utilidad, corrección y credibilidad en la evaluación de documentos del TREC *Health Misinformation Track* 2021, en el apartado 1.2.3 Datos utilizados para este trabajo.

documentos dando lugar a una distribución más dispersa a lo largo de las diferentes puntuaciones, sobre todo de los útiles.

Los documentos perjudiciales se reparten principalmente entre -1 y -2. En cambio, el *score* -3 (documentos incorrectos que son muy creíbles) es el minoritario de las puntuaciones negativas, hecho positivo ya que es el escenario más preocupante.

2.3.2 Estadísticas descriptivas de los documentos evaluados

Distribución general de documentos *helpful/harmful*

Al igual que en 2020, la mayoría de las consultas tienen más documentos útiles que perjudiciales (véase Figura 16). Sin embargo, también como en la edición de 2020, se detectan consultas cuya desinformación supera a la información correcta. Entre las consultas destacan:

- **Topic 101:** *Will wearing an ankle brace help heal achilles tendonitis?* (¿Usar un soporte para el tobillo ayuda a curar la tendinitis de Aquiles?)[20]
- **Topic 102:** *Is a tepid sponge bath a good way to reduce fever in children?* (¿Un baño con esponja tibia es una buena forma de reducir la fiebre en niños?)[21]
- **Topic 103:** *Can folic acid help improve cognition and treat dementia?* (¿Puede el ácido fólico mejorar la cognición y tratar la demencia?)[22]
- **Topic 104:** *Does duct tape work for wart removal?* (¿La cinta adhesiva es eficaz para eliminar verrugas?)[23]
- **Topic 128:** *Does steam from a shower help croup?* (¿El vapor de la ducha ayuda a tratar el crup?) [24]

[20] No existe evidencia concluyente que respalde que el uso de soportes para el tobillo acelere la curación de la tendinitis de Aquiles. Respaldado por la evidencia en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3134723/

[21] Método tradicional que consiste en pasar una esponja mojada en agua templada por el cuerpo para ayudar a bajar la fiebre. No se ha establecido su eficacia para reducir la fiebre. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/9115527/

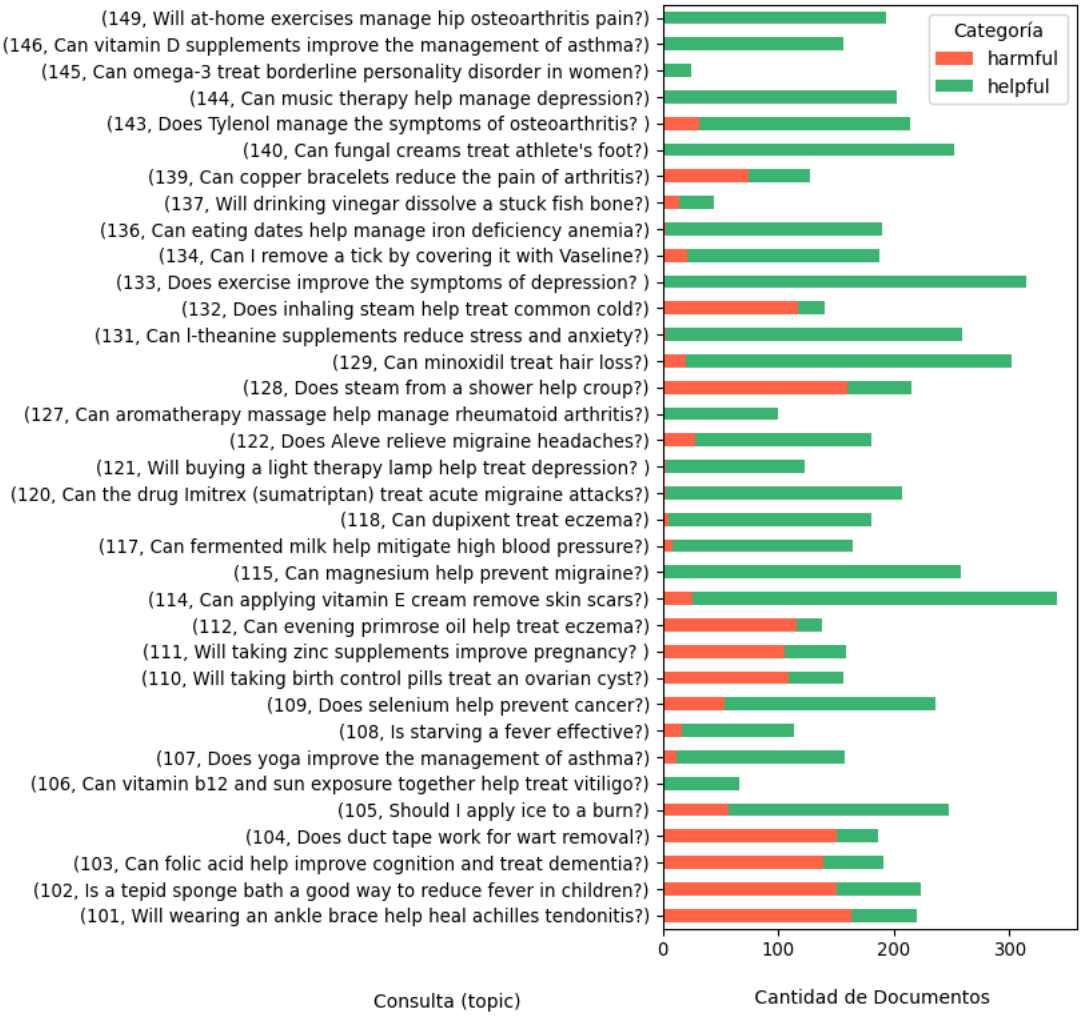
[22] Algunos estudios han explorado el efecto del ácido fólico en la función cognitiva, pero los resultados son limitados y no concluyentes para el tratamiento de la demencia. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/14584018/

[23] No existe evidencia para el uso de cinta adhesiva para eliminar verrugas. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/22972052/

[24] El crup es una infección respiratoria infantil que provoca inflamación en la laringe y la tráquea, causando tos perruna y dificultad para respirar. No se ha demostrado que el vapor de la ducha tenga un efecto terapéutico claro sobre el crup en niños. Respaldado por la evidencia en https://www.webmd.com/children/news/20060314/humidity-may-not-help-kids-with-croup

Observamos como se tratan, en su mayoría, de tratamientos caseros, que siguen siendo más susceptibles de generar resultados de baja calidad o erróneos.

Figura 16. Distribución de documentos útiles (*helpful*) y perjudiciales (*harmful*) por consulta en 2021.[25]



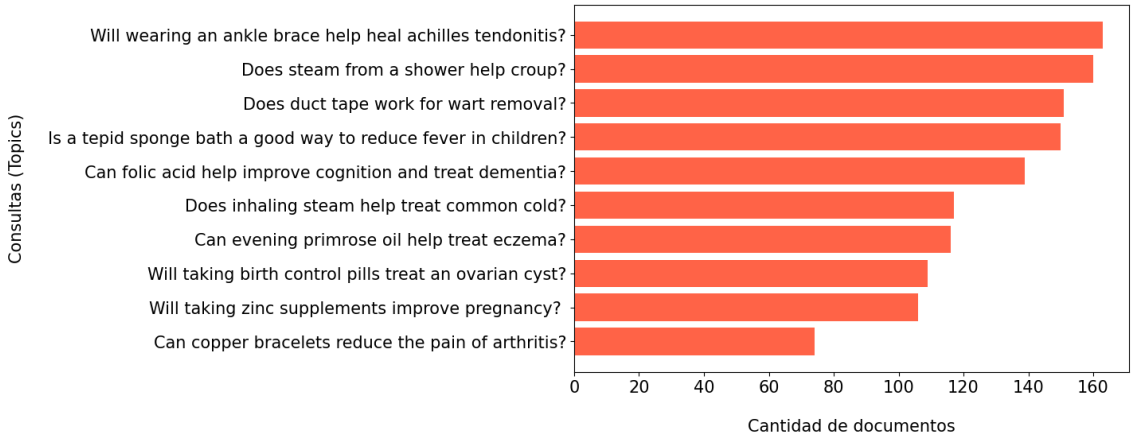
Fuente: Elaboración propia.

[25] Se observa que, aunque inicialmente se definen 50 consultas (*topics)* para esta edición, al concatenar con los documentos (*qrels)* para hacer el recuento, los *topics* 113, 116, 119, 123, 124, 125, 126, 130, 135, 138, 141, 142, 147, 148 y 150 desaparecen. Esto es debido a que no existen documentos para esas consulta en los ficheros de *Qrels 2021*.

Consultas con más documentos perjudiciales (*harmful*)

La Figura 17 muestra las diez consultas con mayor volumen de documentos perjudiciales en 2021.

Figura 17. Top 10 de consultas con mayor número de documentos perjudiciales en 2021.



Fuente: Elaboración propia.

Además de las consultas comentadas en el apartado anterior, se incluyen otras cinco consultas con una gran cantidad de documentos perjudiciales:

- *Does inhaling steam help treat common cold?* (¿Inhalar vapor ayuda a tratar el resfriado común?)[26]
- *Can evening primrose oil help treat eczema?* (¿El aceite de onagra ayuda a tratar el eccema?)[27]
- *Will taking birth control pills treat an ovarian cyst?* (¿Tomar anticonceptivos trata un quiste ovárico?)[28]

[26] No hay evidencia suficiente que respalde que inhalar vapor alivie los síntomas del resfriado común de forma eficaz. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/28849871/

[27] El aceite de onagra es un aceite vegetal obtenido de las semillas de la planta Oenothera biennis. No se ha demostrado beneficios clínicos consistentes en su uso para tratar el eccema. Respaldado por la evidencia en https://pmc.ncbi.nlm.nih.gov/articles/PMC292973/

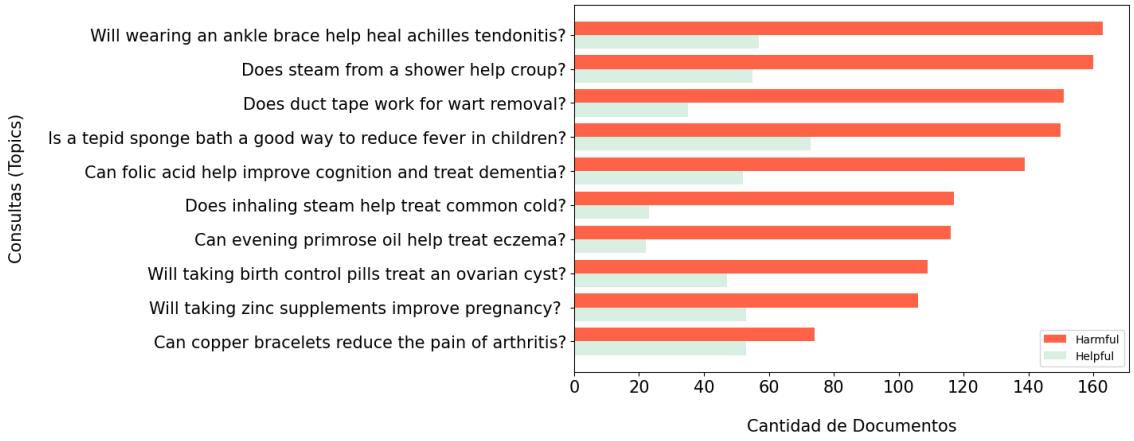
[28] Los anticonceptivos orales pueden utilizarse como tratamiento para algunos tipos de quistes ováricos funcionales. Sin embargo, se recomienda una conducta expectante durante dos o tres ciclos antes de considerar tratamiento quirúrgico. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/24782304/

- *Will taking zinc supplements improve pregnancy?* (¿Tomar suplementos de zinc mejora el embarazo?)[29]
- *Can cooper bracalets recuce the pain of arthritis?* (¿Las pulseras de cobre reducen el dolor causado por la artritis?)[30]

Al comparar estas diez consultas con sus respectivos documentos de tipo *helpful* (véase Figura 18), observamos que para todos los casos los documentos perjudiciales superan a los útiles, hecho que representa un contraste con lo observado en la edición de 2020.

En esa edición, incluso en los temas con más desinformación, la cantidad de documentos útiles eran mayoritarios. En cambio, en 2021, las consultas con más desinformación no cuentan con ese equilibrio, lo que puede afectar negativamente la calidad de la información a la que acceden los usuarios.

Figura 18. Top 10 consultas con más documentos perjudiciales en 2021 y comparación con útiles.



Fuente: Elaboración propia.

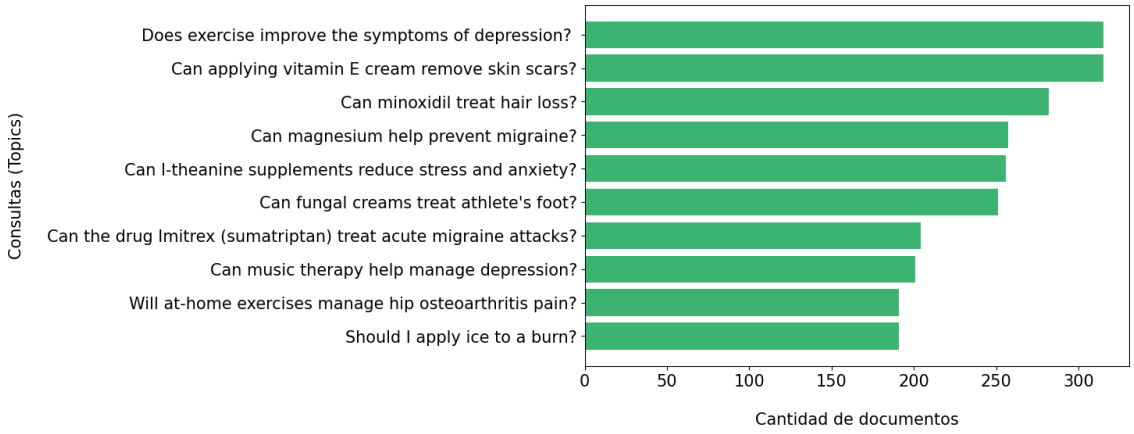
[29] Los suplementos de zinc son productos nutricionales utilizados para mejorar el sistema inmunitario o la fertilidad. Pueden tener efectos positivos en algunos indicadores de salud materna, pero la evidencia no es concluyente para recomendar su uso generalizado. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/33724446/

[30] Las pulseras de cobre se usan en terapias alternativas, asociadas con la creencia de que reduce el dolor articular en enfermedades como la artritis. No han mostrado eficacia superior al placebo en el tratamiento del dolor causado por la artritis. Respaldado por la evidencia en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3774818/

Consultas con más documentos útiles (*helpful*)

Con respecto a las consultas con más documentos útiles etiquetados por los jueces (véase Figura 19), predominan temas más específicos o clínicos, como el uso de *minoxidil* para la caída del cabello, el uso de cremas antifúngicas o fármacos específicos como el *Imitrex*.

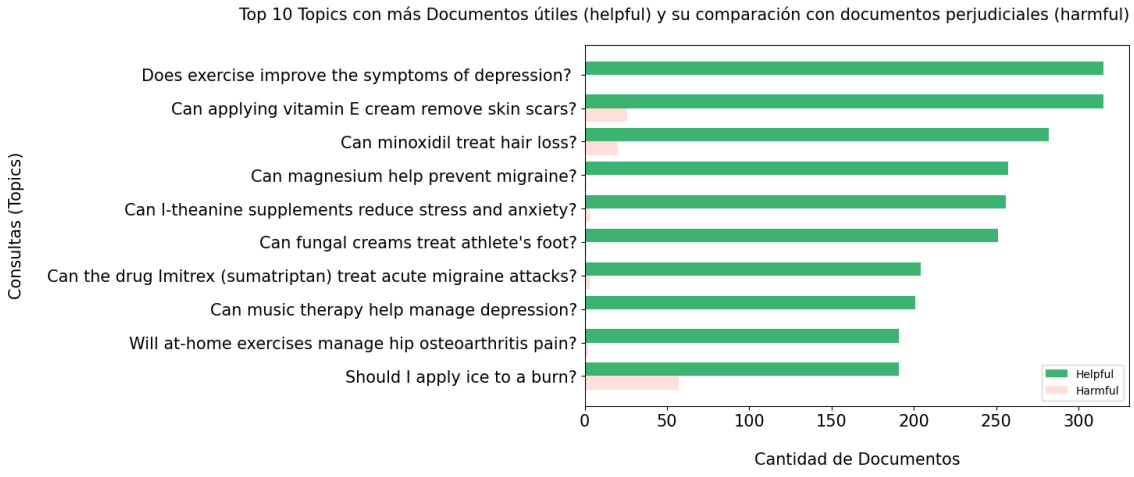
Figura 19. Top 10 de consultas con mayor número de documentos útiles en 2021.



Fuente: Elaboración propia.

Al observar su comparación con los documentos perjudiciales (véase Figura 20), se confirma que, a diferencia de las consultas con mayor presencia de desinformación, estos temas están mucho mejor representados por fuentes de calidad. Tan solo en uno de los casos, la consulta sobre aplicar hielo en quemaduras, los documentos perjudiciales tienen una cantidad considerable, aunque siguen siendo minoría.

Figura 20. Top 10 consultas con más documentos útiles en 2021 y comparación con perjudiciales.



Fuente: Elaboración propia.

2.3.3 Conclusión del análisis de 2021

En general, el análisis de la edición de 2021 confirma lo que ya se había visto en 2020: la mayoría de los documentos son útiles. Sin embargo, este año los documentos perjudiciales han aumentado en proporción y son mayoritarios en aquellas consultas con más desinformación.

Como ya ocurría en la edición anterior, la mayoría de las consultas con un gran volumen de documentos perjudiciales tienen que ver con remedios caseros, prácticas sin base científica o afirmaciones dudosas. En cambio, las consultas con más contenido útil están relacionadas con tratamientos conocidos, mención a fármacos o hábitos saludables.

2.4 Análisis exploratorio de la edición TREC 2022

2.4.1 Cantidad de consultas, documentos y distribución de categorías

Figura 21. Resumen de documentos útiles y perjudiciales en TREC 2020, 2021 y 2022.

```
======================================
Resumen de documentos de 2020:
======================================
Total de documentos: 7256

Distribución por categoría:
         total  cantidad porcentaje
harmful   7256       805      11.1%
helpful   7256      6451      88.9%
```

```
======================================
Resumen de documentos de 2021:
======================================
Total de documentos: 6469

Distribución por categoría:
         total  cantidad porcentaje
harmful   6469      1596      24.7%
helpful   6469      4873      75.3%
```

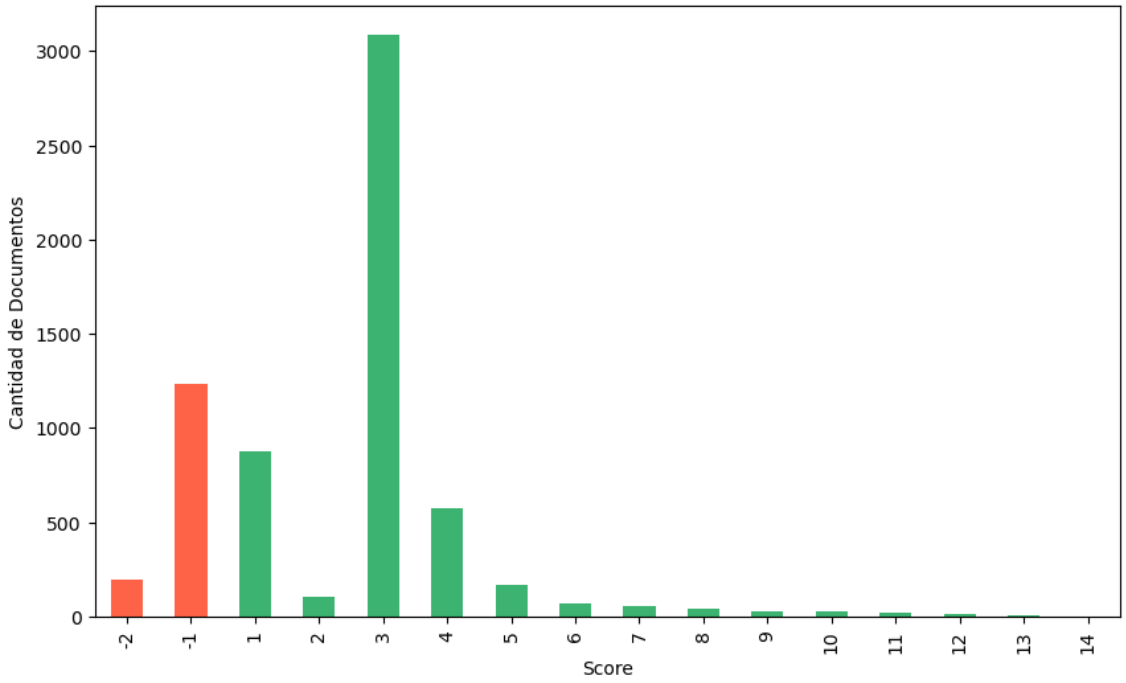
```
======================================
Resumen de documentos de 2022:
======================================
Total de documentos: 6501

Distribución por categoría:
         total  cantidad porcentaje
harmful   6501      1434      22.1%
helpful   6501      5067      77.9%
```

Fuente: Elaboración propia.

En la edición de 2022, se evaluaron un total de 6501 documentos, con una proporción del 77,9 % de documentos útiles y un 22,1 % de perjudiciales. Aunque se mantiene una mayoría en contenido útil, el porcentaje de desinformación es superior al observado en 2020 (11 %) y más cercado al de 2021 (24,7 %).

Con respecto a las puntuaciones, observamos una gran concentración de documentos útiles en el *score 3* y de documentos perjudiciales en el *score -1* (véase Figura 22).

Figura 22. Distribución de documentos por score en TREC 2022.



Fuente: Elaboración propia.

Un aspecto relevante de esta edición es que, aunque la escala era sobre 4, se aplicó una segunda ronda de evaluación para los documentos útiles, que permitió crear una nueva escala sobre 14 basada en juicios de preferencia entre documentos *helpful.*[31]

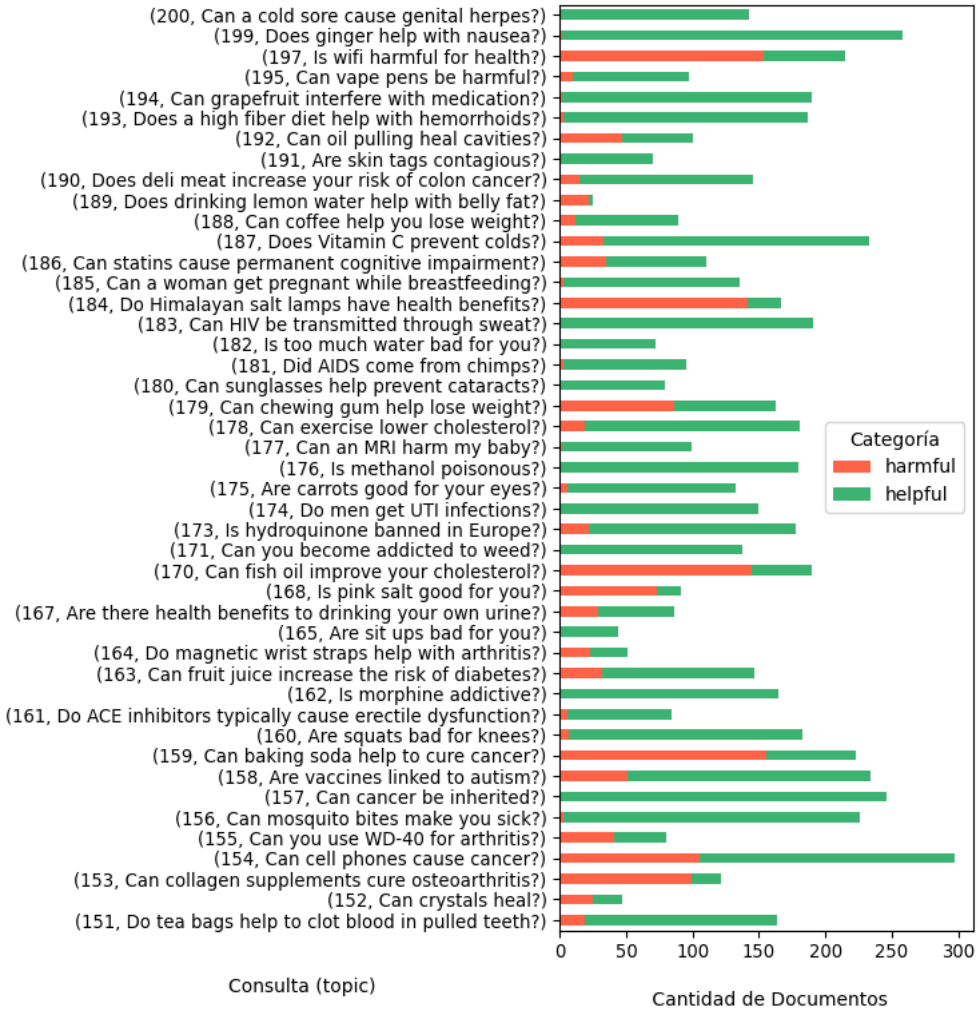
2.4.2 Estadísticas descriptivas de los documentos evaluados

Distribución general de documentos *helpful/harmful*

La Figura 23 muestra la distribución de documentos útiles y perjudiciales en las 50 consultas de 2022. Como en años anteriores, los documentos útiles predominan en la mayoría de las consultas, aunque sigue habiendo algunos temas donde la desinformación es mayoritaria.

[31] Véase Tabla 3: Criterios combinados de utilidad, corrección y credibilidad en la evaluación de documentos del TREC *Health Misinformation Track* 2022, en el apartado 1.2.3 Datos utilizados para este trabajo.

Figura 23. Distribución de documentos útiles (*helpful*) y perjudiciales (*harmful*) por consulta en 2022.[32]



Fuente: Elaboración propia.

Consultas con más documentos perjudiciales (*harmful*)

La Figura 24 muestra las consultas con mayor número de documentos etiquetados como perjudiciales en 2022. Estas consultas están mayoritariamente relacionadas con afirmaciones pseudocientíficas o creencias populares sin respaldo médico, patrón repetido en las ediciones anteriores. Las consultas incluidas son:

[32] Se observa que, aunque inicialmente se definen 50 consultas (*topics)* para esta edición, al concatenar con los documentos (*qrels)* para hacer el recuento, los *topics* 166, 169, 172, 196 y 198 desaparecen. Esto es debido a que no existen documentos para esas consultas en los ficheros de *Qrels 2022*.

- Can baking soda help to cure cancer? (¿El bicarbonato de sodio ayuda a curar el cáncer?)[33]
- Is wifi harmful for health? (¿El wifi es perjudicial para la salud?)[34]
- Can fish oil improve your cholesterol? (¿El aceite de pescado mejora el colesterol?)[35]
- Do Himalayan salt lamps have health benefits? (¿Las lámparas de sal del Himalaya tienen beneficios para la salud?)[36]
- Can cell phones cause cancer? (¿Los teléfonos móviles causan cáncer?)[37]
- Can collagen supplements cure osteoarthritis? (¿Los suplementos de colágeno curan la artrosis?)[38]
- Can chewing gum help lose weight? (¿Masticar chicle ayuda a perder peso?)[39]
- Is pink salt good for you? (¿La sal rosa es buena para la salud?)[40]
- Are vaccines linked to autism? (¿Las vacunas están relacionadas con el autismo?)[41]
- Can oil pulling heal cavities? (¿El enjuague con aceite cura las caries?)[42]

[33] El bicarbonato de sodio es una sustancia alcalina que se ha utilizado como remedio casero, pero no tiene propiedades anticancerígenas demostradas. Respaldado por la evidencia en https://www.webmd.com/a-to-z-guides/baking-soda-do-dont

[34] El WiFi mite ondas de radio no ionizantes, similares a las de un microondas pero en potencias mucho menores, y no se ha demostrado que afecten negativamente a la salud humana. Respaldado por la evidencia en https://www.canada.ca/en/health-canada/services/health-risks-safety/radiation/everyday-things-emit-radiation/wi-fi.html

[35] El aceite de pescado contiene ácidos grasos omega-3, que pueden tener beneficios cardiovasculares, pero no está demostrado que mejoren directamente los niveles de colesterol en todos los casos. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/18774613/

[36] Las lámparas de sal de Himalaya son bloques de sal que se iluminan desde dentro. Se les atribuyen propiedades sanadoras sin evidencia científica. Respaldado por la evidencia en https://www.webmd.com/balance/himalayan-salt-lamps

[37] Los teléfonos móviles emiten radiación no ionizante, y no hay evidencia concluyente que los relacione directamente con el desarrollo de cáncer. Respaldado por la evidencia en https://www.cancer.gov/about-cancer/causes-prevention/risk/radiation/cell-phones-fact-sheet

[38] La Osteoartritis es una enfermedad degenerativa de las articulaciones. Aunque el colágeno puede mejorar la salud articular en algunos casos, no es una cura. Respaldado por la evidencia en https://link.springer.com/article/10.1007/s40744-020-00240-5

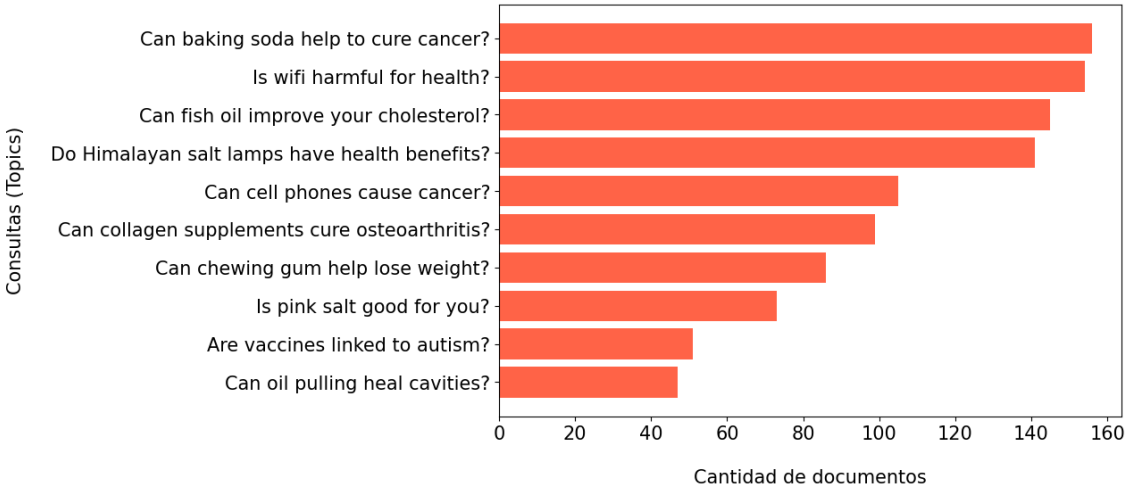
[39] Masticar chicle puede contribuir a la sensación de saciedad, pero no hay evidencia sólida que lo vincule con una pérdida de peso significativa. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/22076595/

[40] La sal rosa contiene trazas de minerales, pero no tiene beneficios demostrados superiores a la sal común. Respaldado por la evidencia en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7603209/

[41] La relación entre vacunas y autismo ha sido desmentida en múltiples estudios científicos rigurosos. Respaldado por la evidencia en https://pubmed.ncbi.nlm.nih.gov/24814559/

[42] El enjuague con aceite es una práctica ayurvédica tradicional que consiste en enjuagarse la boca con aceite. No hay evidencia de que cure las caries. Respaldado por la evidencia en https://www.mouthhealthy.org/en/az-topics/o/oil-pulling

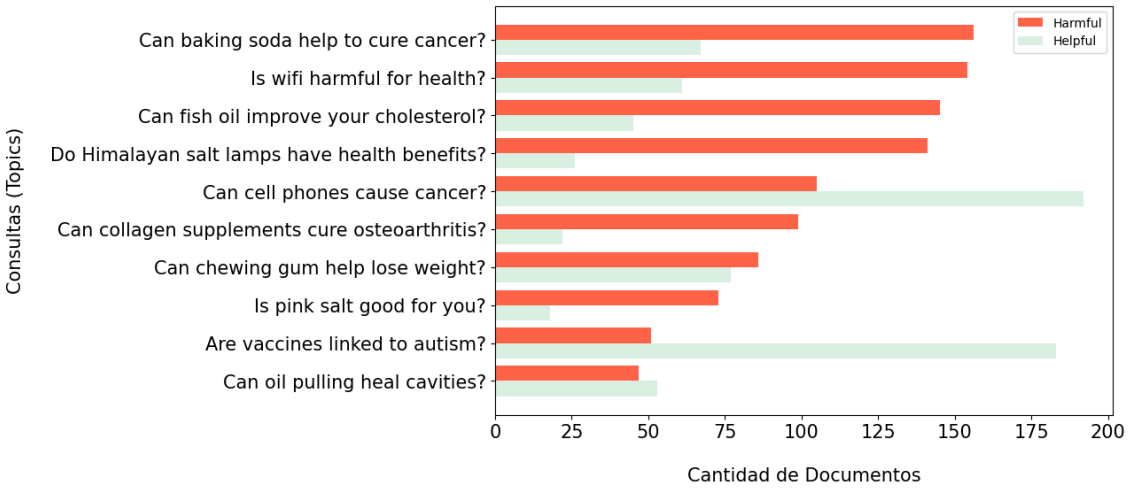
Figura 24. Top 10 de consultas con mayor número de documentos perjudiciales en 2022



Fuente: Elaboración propia.

Como se observa en la Figura 25, en algunas de estas consultas el número de documentos perjudiciales iguala o incluso supera a los documentos útiles. Sin embargo, a diferencia de lo observado en 2021, este año el contenido útil logra mantener una presencia más equilibrada en algunos de los temas, superando o igualando a la cantidad de desinformación para ciertas consultas.

Figura 25. Top 10 consultas con más documentos perjudiciales en 2022 y comparación con útiles.

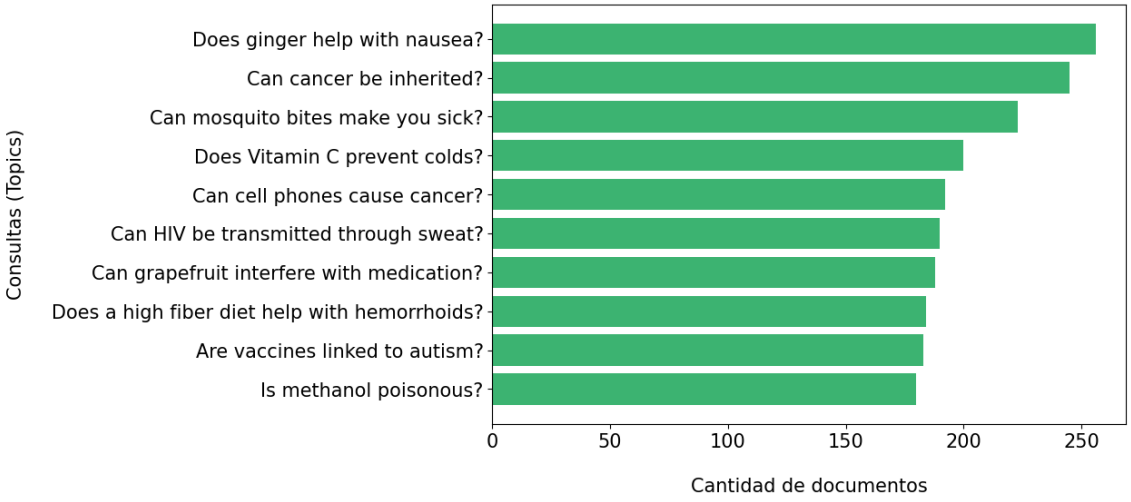


Fuente: Elaboración propia

Consultas con más documentos útiles (*helpful*)

En la Figura 26 se muestran las diez consultas con más documentos útiles de la edición de 2022 y en la Figura 27, su comparativa con la cantidad de documentos perjudiciales equivalentes.

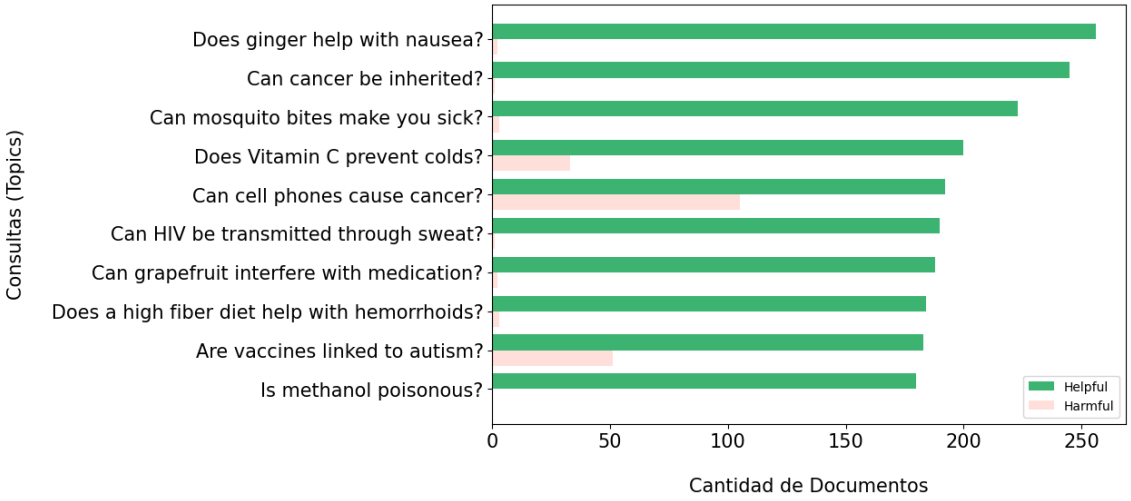
Figura 26. Top 10 de consultas con mayor número de documentos útiles en 2022



Fuente: Elaboración propia.

La mayoría de estas consultas presentan una proporción dominante de documentos útiles y una cantidad muy baja de desinformación. Solo algunos casos como la consulta sobre teléfonos móviles, las vacunas y la Vitamina C atraen documentos perjudiciales.

Figura 27. Top 10 consultas con más documentos útiles en 2022 y comparación con perjudiciales.



Fuente: Elaboración propia.

2.4.3 Conclusión del análisis de 2022

Se vuelve a observar que las consultas con más desinformación suelen girar en torno a remedios caseros, afirmaciones virales o temas pseudocientíficos (como el bicarbonato, el wifi o las lámparas de sal). Por otro lado, las consultas con mayor número de documentos útiles tienden a referirse a preguntas clínicas más concretas.

3 Desarrollo de predictores de desinformación médica

Una vez realizado el análisis exploratorio y conocida la distribución de documentos útiles y perjudiciales por consulta, el siguiente paso consiste en desarrollar predictores que, dada una consulta médica que una persona podría introducir en un buscador, sean capaces de anticipar si recuperará mayoritariamente contenido útil o, por el contrario, perjudicial.

El objetivo es aproximarse a una estimación automatizada del riesgo de desinformación asociado a cada consulta, evaluando si existen patrones que nos permita identificar consultas potencialmente perjudiciales antes incluso de que se realice la búsqueda.

Para ello, se han explorado tres vías distintas de predicción:

1. Un análisis de frecuencia léxica, para estudiar si las consultas formuladas con palabras más comunes o técnicas influyen en el tipo de documentos recuperados.
2. Un clasificador de sesgo, entrenado para detectar si una consulta está redactada de una forma que pueda influir en los resultados.
3. Por último, se usará una estrategia basada en Modelos Grandes de Lenguaje (MGL) usando *prompting*[43] para obtener su estimación.

3.1 Análisis de la frecuencia léxica en las consultas

Este capítulo tiene como objetivo analizar si las características lingüísticas de las preguntas, como la frecuencia de las palabras utilizadas, tienen alguna relación con la calidad de los documentos que recuperan.

3.1.1 Introducción

3.1.1.1 Motivación del análisis

Como vimos en el análisis del capítulo anterior, una de las hipótesis planteadas es que quizás las consultas con terminología médica más técnica (por ejemplo, tratamientos o nombres

[43] Un prompt es una instrucción o conjunto de indicaciones que se le da a un modelo de lenguaje para que genere una respuesta concreta.

de fármacos) devuelvan más documentos útiles. En cambio, las consultas con palabras más comunes y coloquiales pueden estar más relacionadas con documentos perjudiciales.

3.1.1.2 Librería wordfreq y preparación de los datos

Para explorar esta idea, se ha realizado un análisis de la frecuencia de las palabras que aparecen en las consultas médicas utilizando la librería **wordfreq** de Python (véase Figura 28).

Figura 28. Código Python para el cálculo de frecuencia media y mínima.

```python
# Frecuencia media de palabras por oración
def obtener_frecuencia_media(sentence, lang='en'):
    if not isinstance(sentence, str):
        return 0.0
    words = sentence.lower().split()
    words = [word.strip(string.punctuation) for word in words]
    words = [word for word in words if word.isalpha() and word not in stop_words]
    freqs = [word_frequency(word, lang, wordlist='best') for word in words]
    if not freqs:
        return 0.0
    return sum(freqs) / len(freqs)

# Frecuencia mínima de palabras por oración
def obtener_frecuencia_minima(sentence, lang='en'):
    if not isinstance(sentence, str):
        return 0.0
    words = sentence.lower().split()
    words = [word.strip(string.punctuation) for word in words]
    words = [word for word in words if word.isalpha() and word not in stop_words]
    freqs = [word_frequency(word, lang, wordlist='best') for word in words]
    freqs = [f for f in freqs if f > 0]
    if not freqs:
        return 0.0
    return min(freqs)

# Añadir columna con frecuencia media y mínima
qrels_2020_freq['mean_freq'] = qrels_2020_freq['question'].apply(obtener_frecuencia_media)
qrels_2020_freq['min_freq'] = qrels_2020_freq['question'].apply(obtener_frecuencia_minima)
```

Python

Fuente: Elaboración propia

Esta herramienta permite saber con qué frecuencia se usan ciertas palabras en el lenguaje cotidiano, basándose en grandes colecciones de textos en varios idiomas. Gracias a esto, podemos identificar si una consulta está formulada con palabras muy comunes o poco frecuentes, lo cual podría influir en los resultados que devuelve el sistema de búsqueda.

Se aplicó esta herramienta a todas las consultas médicas de las ediciones de TREC 2020, 2021 y 2022. Para cada consulta se calcularon dos métricas:

- **La frecuencia media,** es decir, el promedio de frecuencia de todas las palabras que la componen.
- **La frecuencia mínima,** que representa la palabra menos frecuente dentro de la consulta.

3.1.1.3 Correlaciones de Pearson y Spearman

El último paso del análisis de las frecuencias de las palabras con respecto a la cantidad de documentos recuperados será analizar la correlación. Para ello es importante explicar qué significa una correlación en el contexto de este análisis.

Una **correlación** es una medida estadística que permite saber si existe una relación entre dos variables y en qué dirección se da esa relación. Su valor puede oscilar entre **-1 y 1**:

- Un valor cercano a 1 indica una correlación positiva fuerte: cuando una variable aumenta, la otra también lo hace.
- Un valor cercano a -1 indica una correlación negativa fuerte: cuando una variable sube, la otra tiende a bajar.
- Un valor cercano a 0 indica poca o ninguna relación entre las variables.

En este caso, se busca comprobar si existe relación entre la frecuencia de las palabras en una consulta (ya sea la media o la mínima) y el número de documentos útiles o perjudiciales que esa consulta recupera. Para ello se calcularon dos tipos de correlaciones:

- **Pearson:** mide la relación lineal entre dos variables. Es útil cuando se espera que los cambios en una variable estén acompañados por cambios proporcionales en la otra. Sin embargo, es sensible a valores extremos (*outliers*).
- **Spearman:** mide la relación monótona entre dos variables. Es más robusta ante datos no distribuidos normalmente o con relaciones no lineales, y resulta ideal para detectar tendencias más generales.

La Figura 29 muestra el código empleado para calcular las correlaciones de Pearson y Spearman.

Figura 29. Cálculo de correlaciones entre frecuencia léxica y categoría de documento (2020)

```python
# Crear un DataFrame con las correlaciones
correlation_data = {
    'mean_freq': {
        'harmful (Pearson)': pearsonr(mean_freq_2020['mean_freq'], mean_freq_2020['harmful'])[0],
        'harmful (Spearman)': spearmanr(mean_freq_2020['mean_freq'], mean_freq_2020['harmful'])[0],
        'helpful (Pearson)': pearsonr(mean_freq_2020['mean_freq'], mean_freq_2020['helpful'])[0],
        'helpful (Spearman)': spearmanr(mean_freq_2020['mean_freq'], mean_freq_2020['helpful'])[0],
    },
    'min_freq': {
        'harmful (Pearson)': pearsonr(min_freq_2020['min_freq'], min_freq_2020['harmful'])[0],
        'harmful (Spearman)': spearmanr(min_freq_2020['min_freq'], min_freq_2020['harmful'])[0],
        'helpful (Pearson)': pearsonr(min_freq_2020['min_freq'], min_freq_2020['helpful'])[0],
        'helpful (Spearman)': spearmanr(min_freq_2020['min_freq'], min_freq_2020['helpful'])[0],
    }
}

correlation_df = pd.DataFrame(correlation_data)

# Heatmap
plt.figure(figsize=(8, 5))
sns.heatmap(correlation_df, annot=True, cmap='RdBu_r', center=0, linewidths=0.5, fmt=".4f")
plt.title("Correlaciones entre frecuencia (media y mínima) y categoría en 2020")
plt.tight_layout()
plt.show()
```

Python

Fuente: Elaboración propia.

3.1.2 Análisis de frecuencia para el año 2020

3.1.2.1 Distribución de frecuencia media y mínima y tipo de documento

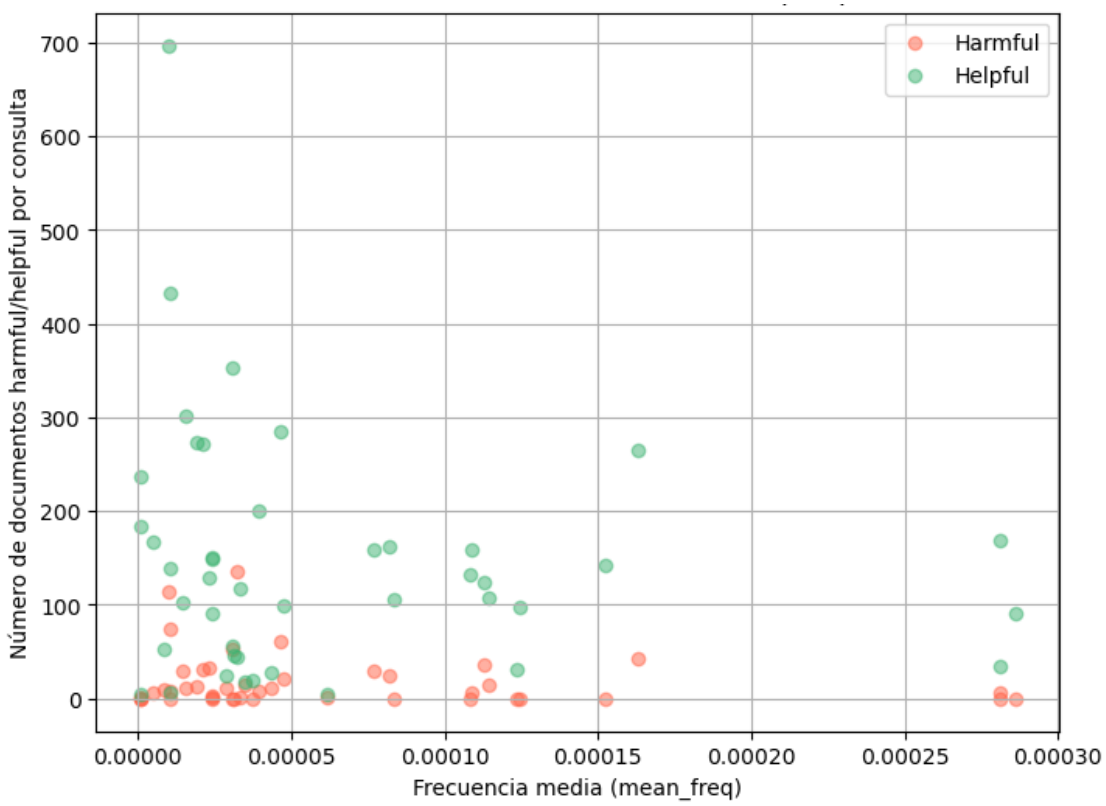
Después de calcular la frecuencia media y mínima de las palabras que componen cada consulta, se quiso analizar si había algún tipo de relación entre el tipo de lenguaje utilizado y la cantidad de documentos de cada tipo que tiene asociada esa consulta.

Frecuencia media

La Figura 30 muestra un gráfico de dispersión en el que podemos observar que tanto los documentos útiles como los perjudiciales tienden a concentrarse en frecuencias bajas, lo que indica que ambos tipos de consultas suelen estar asociados a términos poco comunes en el idioma. Aun así, el volumen de documentos útiles es superior al de los perjudiciales.

Además, se aprecia una ligera tendencia negativa a medida que aumenta la frecuencia media de las palabras en una consulta. Esto refuerza la idea de que cuanto más común es el lenguaje, menos documentos útiles se recuperan.

Figura 30. Frecuencia media vs. número de documentos *helpful/harmful* por consulta (2020).

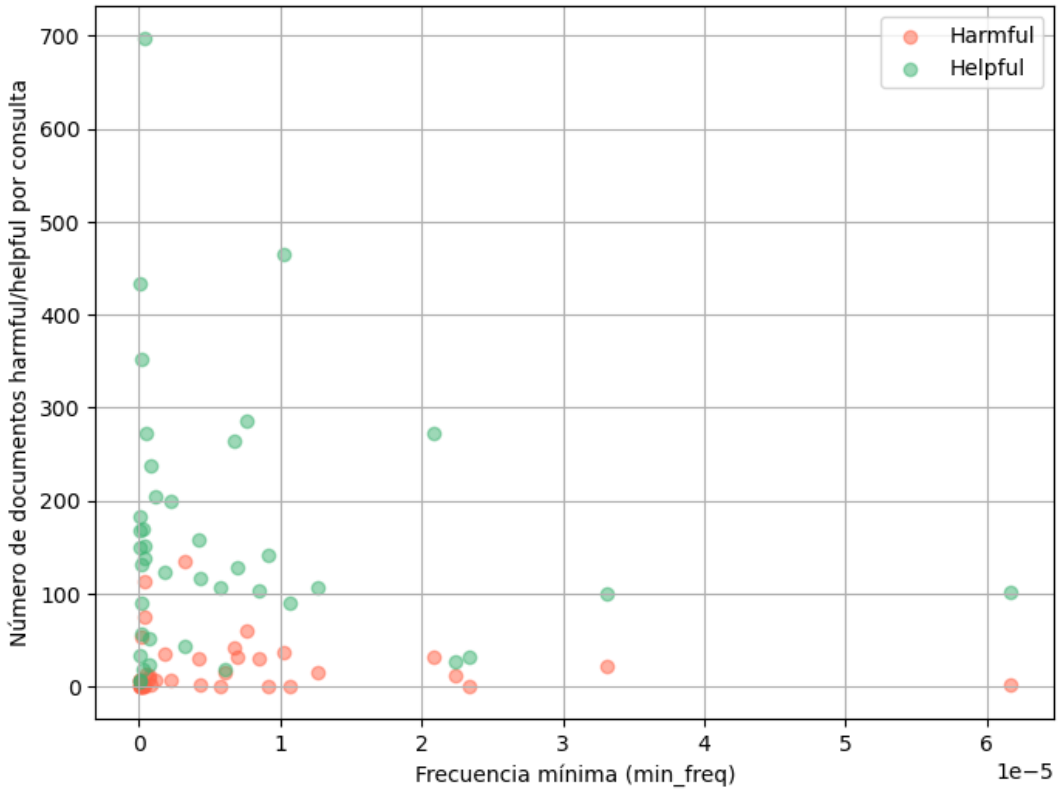


Fuente: Elaboración propia.

Frecuencia mínima

En este caso observamos un comportamiento similar al de la frecuencia media, pero aún más marcado en el caso de los documentos útiles que se concentran en su mayoría en frecuencias mínimas muy bajas y con un volumen de documentos que supera con creces a los documentos perjudiciales, que se encuentran más dispersos entre las frecuencias mínimas bajas (véase Figura 31).

Figura 31. Frecuencia mínima vs. número de documentos *helpful/harmful* por consulta (2020).

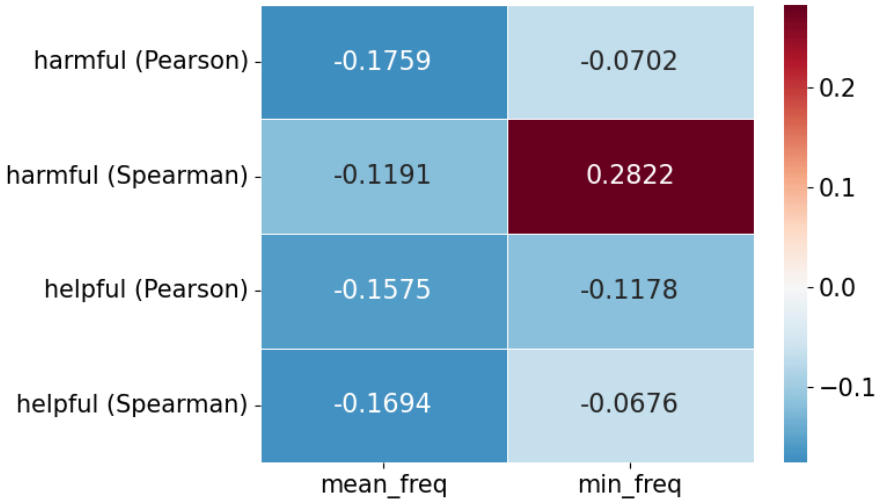


Fuente: Elaboración propia.

3.1.2.2 Correlaciones entre frecuencia y tipo de documento

Para completar el análisis se calcularon las correlaciones de Pearson y de Spearman entre las frecuencias de las palabras (tanto media como mínima) y el número de documentos recuperados de cada tipo (véase Figura 32).

Figura 32. Correlaciones entre frecuencia léxica y número de documentos *helpful/harmful* (2020).



Fuente: Elaboración propia.

Frecuencia media

Las correlaciones con la frecuencia media (mean_freq) son negativas y bajas tanto para documentos útiles como perjudiciales. Este patrón simétrico implica que la frecuencia media no sirve como discriminador fiable entre ambos tipos de documentos, ya que ambos tienden a concentrarse en consultas con términos poco frecuentes.

Frecuencia mínima

En cuanto a la frecuencia mínima (*min_freq*), los valores también son negativos y cercanos a cero, con la excepción de la correlación Spearman con documentos perjudiciales, en el cual se obtiene una correlación positiva y con un valor superior a los demás (0.2822).

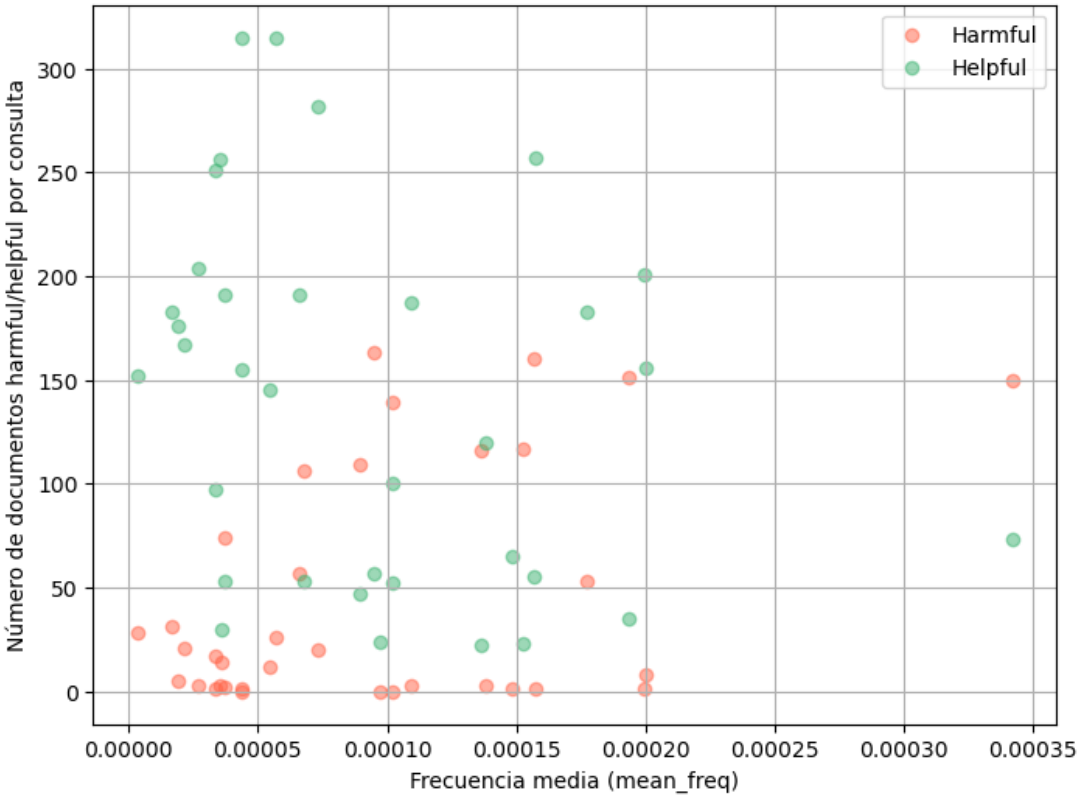
Este valor sugiere que, a mayor frecuencia mínima, es decir, cuando la palabra menos frecuente de la consulta es bastante común en el idioma hay una mayor probabilidad de recuperar documentos perjudiciales.

3.1.3 Análisis de frecuencia para el año 2021

3.1.3.1 Distribución de frecuencia media y mínima y tipo de documento

En el gráfico de dispersión de **frecuencia media** (Figura 33), vemos que los documentos útiles siguen siendo más numerosos, sobre todo cuando las consultas están formuladas con palabras poco frecuentes. Sin embargo, también hay un aumento de volumen de documentos perjudiciales en ese mismo rango de palabras poco comunes, lo cual reduce la distinción que habíamos visto el año anterior.

Figura 33. Frecuencia media vs. número de documentos *helpful/harmful* por consulta (2021).



Fuente: Elaboración propia.

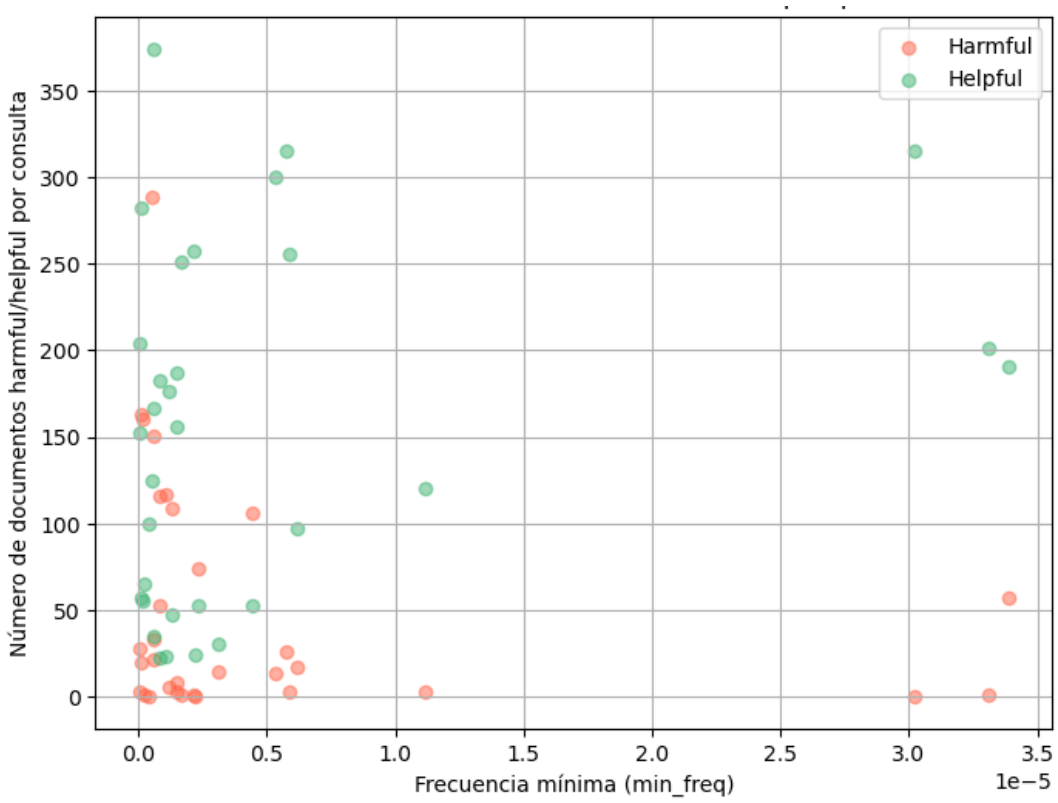
Además, para ambos casos podemos observar mayor dispersión: ya no solo se concentran en rangos de frecuencias de palabras poco comunes, sino que ahora se distribuyen en zonas de palabras más comunes también. Esta mayor variedad sugiere que el tipo de lenguaje de las consultas está más repartido entre categorías, y que el vocabulario por sí solo ya no es un indicador tan claro del tipo de contenido recuperado.

Cabe recordar que, en 2020, las consultas trataban solo sobre COVID-19, lo que explicaba un lenguaje más técnico. En 2021, la mayor variedad de temas generó mayor diversidad léxica.

Aun así, a simple vista parece mantenerse una tendencia positiva en documentos perjudiciales y una tendencia negativa en documentos útiles.

En el gráfico de **frecuencia mínima** (véase Figura 34), parece haber un cambio en la tendencia. Se observa una tendencia negativa en documentos perjudiciales, lo que sugiere que muchas de las consultas asociadas a desinformación contienen al menos una palabra poco común. Esto encaja con la idea de que ciertos documentos perjudiciales usan términos técnicos o específicos, probablemente para parecer más creíbles.

Figura 34. Frecuencia mínima vs. número de documentos *helpful/harmful* por consulta (2021).



Fuente: Elaboración propia.

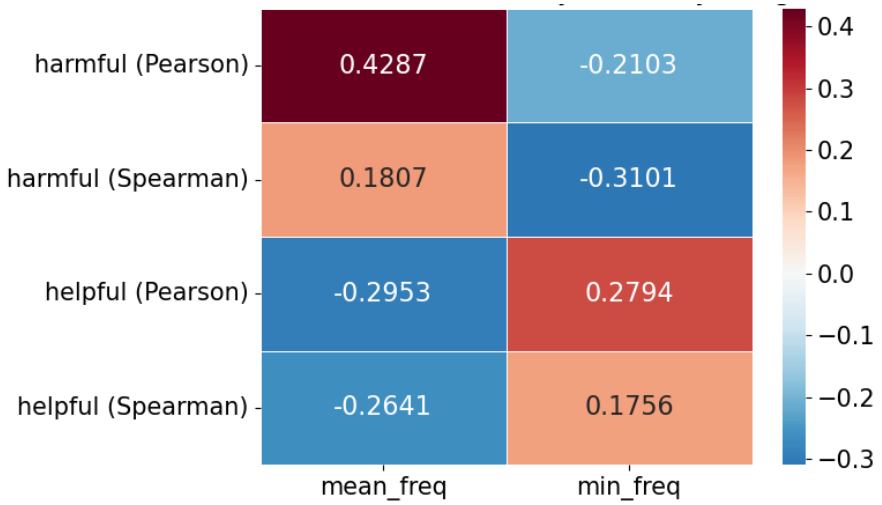
En el caso de los documentos útiles, la mayor parte de las consultas con alto volumen de documentos útiles se concentran en frecuencias mínimas bajas, sin que exista una tendencia evidente. Sin embargo, se observa como en esta edición hay un par de consultas ubicadas en valores de frecuencia mínima muy baja que concentran un volumen excepcionalmente alto de documentos útiles.

Esto nos indica cierta tendencia a la elaboración de documentos dañinos con un lenguaje más profesionalizado o técnico, posiblemente para ganar credibilidad o parecer más fiables a los ojos del lector. Por lo tanto, cuando una consulta incluye términos poco comunes, es más probable que recupere también ese tipo de contenido.

3.1.3.2 Correlaciones entre frecuencia y tipo de documento

Los resultados obtenidos en 2021 muestran patrones diferentes con respecto a lo observado en 2020 (véase Figura 35).

Figura 35. Correlaciones entre frecuencia léxica y número de documentos *helpful/harmful* (2021).



Fuente: Elaboración propia.

Frecuencia media

La correlación entre frecuencia media y documentos perjudiciales es positiva y más fuerte que el año anterior (Pearson: 0.4287 frente a -0.1759 en 2020), lo que refuerza la tendencia positiva que se observó en el gráfico de dispersión. Esto indica que, a diferencia del año anterior, las consultas con palabras más comunes tienden ahora a recuperar más documentos perjudiciales. Esta tendencia se mantiene, aunque con menor intensidad, en la correlación de Spearman, lo que nos indica que la relación tiende a ser lineal.

En el caso de los documentos útiles, la frecuencia media sigue mostrando una correlación negativa, como ya ocurría en 2020, aunque ahora es un poco más marcada. Esto confirma que las preguntas formuladas con palabras más técnicas o especializadas tienden a recuperar documentos más útiles, manteniendo el patrón observado el año anterior.

Frecuencia mínima

Con respecto a la frecuencia mínima, los valores cambian notablemente. Mientras que en 2020 había una correlación positiva entre frecuencia mínima y *harmful* (Spearman: 0.2822), en 2021 esta relación pasa a ser negativa (Spearman: -0.3101). Esto implica que ahora es más probable encontrar documentos dañinos en consultas que contienen al menos una palabra muy poco común, lo que confirma lo que ya vimos en el gráfico de dispersión.

Por otro lado, la correlación entre frecuencia mínima y documentos útiles se vuelve positiva (Pearson: 0.2794), aunque de forma moderada.

3.1.3.3 Conclusiones

En resumen, los datos de 2021 reflejan una dispersión mayor y patrones diferentes. A pesar de que, en el año anterior, la frecuencia mínima servía como buen indicador por sí sola para detectar la desinformación, en esta edición la correlación cambia de signo, lo que nos hace ser escépticos sobre su capacidad predictora. Esto sugiere que su comportamiento no es consistente entre años, y por tanto no puede considerarse una métrica fiable de forma aislada.

En cambio, la frecuencia media, que en 2020 apenas ofrecía información útil, sí muestra una señal clara en 2021: cuanto más comunes son las palabras de una consulta, mayor es la probabilidad de recuperar documentos perjudiciales. Esto confirma nuestra hipótesis inicial de que las consultas formuladas con lenguaje más técnico o especializado tienden a recuperar contenido más útil.

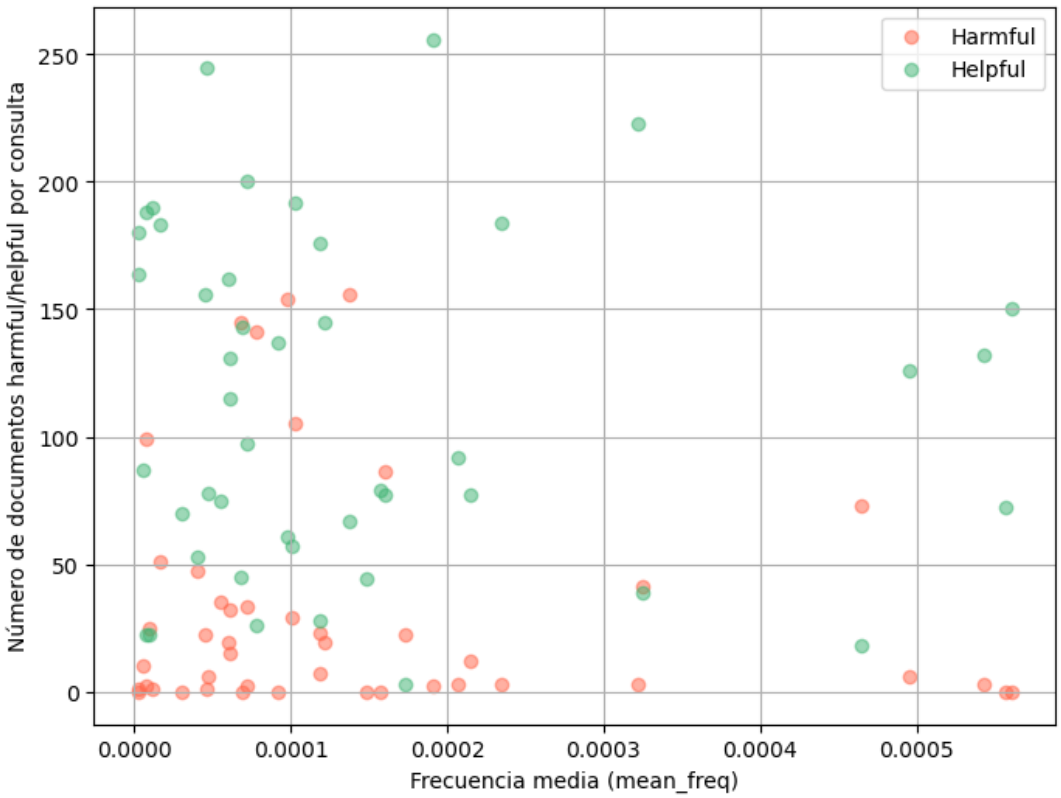
3.1.4 Análisis de frecuencia para el año 2022

3.1.4.1 Distribución de frecuencia media y mínima y tipo de documento

En 2022, igual que en los años anteriores, la mayoría de las consultas se concentran en **frecuencias medias** bajas tanto para documentos útiles como dañinos.

Este año, la separación entre ambos tipos de documentos se ve más difuso ya que ambos siguen una tendencia ligeramente negativa (véase Figura 36).

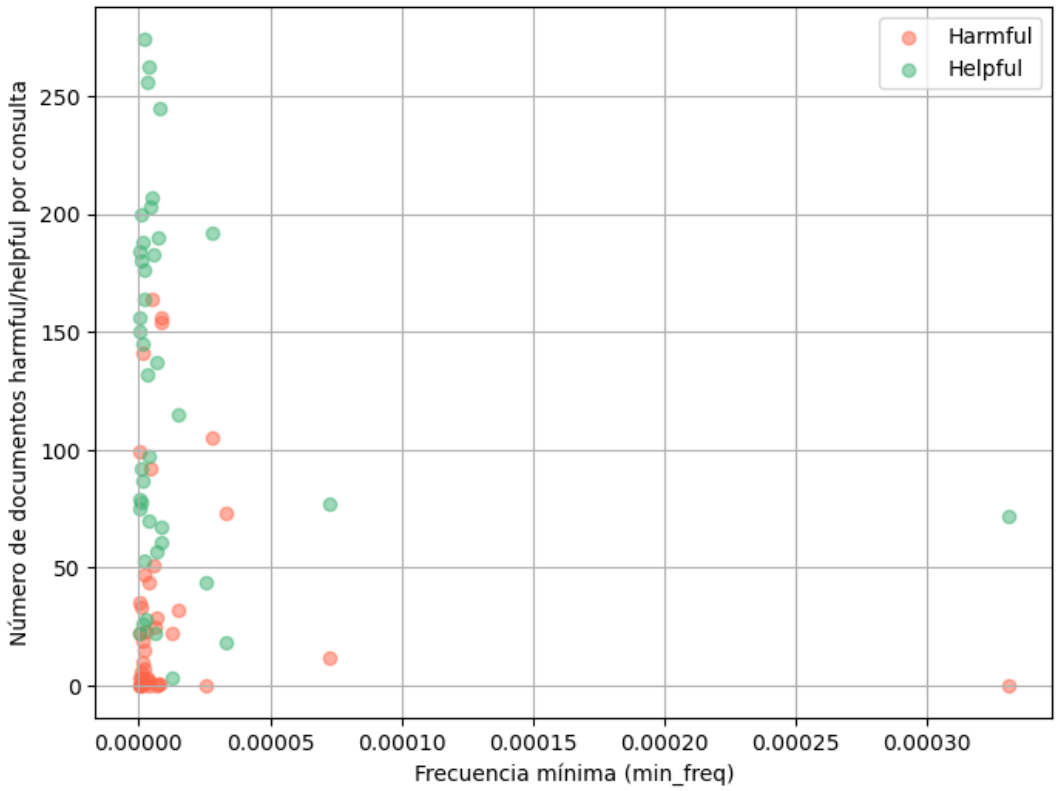
Figura 36. Frecuencia media vs. número de documentos *helpful/harmful* por consulta (2022).



Fuente: Elaboración propia.

En cuanto a la **frecuencia mínima** (véase Figura 37), la tendencia es similar a la observada en 2021. Muchos documentos perjudiciales siguen asociados a consultas que contienen al menos una palabra poco común, pero no se aprecia una diferenciación clara entre consultas que recuperan más documentos útiles o perjudiciales.

Figura 37. Frecuencia mínima vs. número de documentos *helpful/harmful* por consulta (2022).



Fuente: Elaboración propia.

3.1.4.2 Correlaciones entre frecuencia y tipo de documento

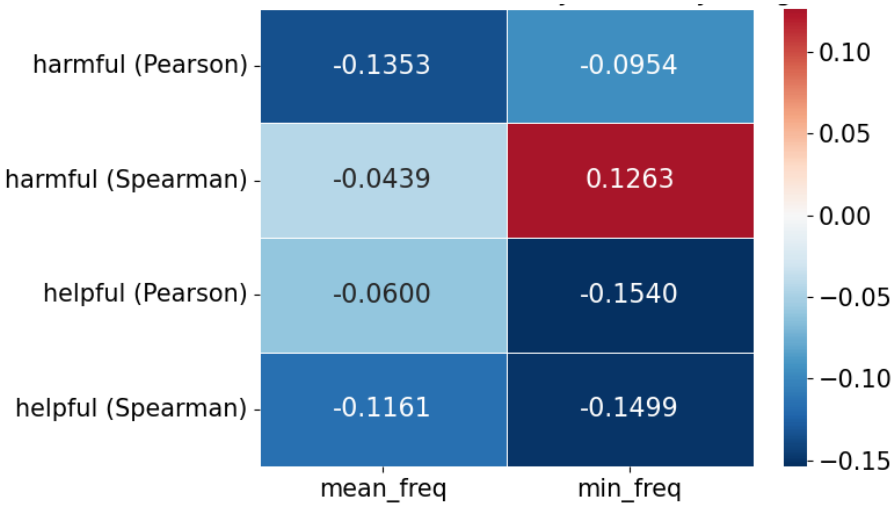
Las correlaciones en 2022 son prácticamente nulas (véase Figura 38), lo que indica muy poca relación entre la frecuencia de las palabras y el número de documentos.

Esto dificulta la identificación automática basada exclusivamente en frecuencia léxica, especialmente en esta edición, donde ambas variables (frecuencia media y frecuencia mínima) pierden su utilidad como predictoras fiables.

En cuanto a la **frecuencia media**, vuelve a observarse lo mismo que ya ocurría en 2020: no hay distinción clara entre consultas que recuperan documentos útiles y aquellas que traen documentos perjudiciales.

En cambio, con la **frecuencia mínima** se refuerza el patrón visto en 2020: cuando la palabra menos frecuente de la consulta es bastante común, hay una mayor probabilidad de recuperar documentos perjudiciales y menos probabilidad de obtener contenido útil.

Figura 38. Correlaciones entre frecuencia léxica y número de documentos *helpful/harmful* (2022).



Fuente: Elaboración propia.

3.1.4.3 Conclusiones

A lo largo de los tres años analizados, se ha observado que las consultas con palabras menos frecuentes suelen recuperar más documentos útiles), mientras que aquellas con lenguaje más común tienden a estar más asociadas a desinformación, especialmente en 2021.

Sin embargo, estos patrones no son consistentes ni lo suficientemente fuertes como para considerarlos buenos predictores por sí solos. Aunque la frecuencia media y mínima aportan información interesante sobre el tipo de lenguaje de las consultas, no permiten anticipar con fiabilidad el tipo de documentos que se van a recuperar.

3.2 Análisis de sesgo

3.2.1 Introducción y modelo utilizado

El objetivo de este capítulo es explorar si el sesgo presente en las consultas médicas puede estar relacionado con el tipo de documentos que se recuperan, especialmente en lo que respecta a contenido perjudicial o útil. La hipótesis es que una consulta formulada de forma sesgada podría tener más probabilidad de generar resultados problemáticos o poco fiables.

Para analizar esta cuestión, se utilizará una arquitectura de referencia en el ámbito del Procesamiento del Lenguaje Natural (PLN): los modelos *Transformer*. Esta arquitectura revolucionó el campo de la inteligencia artificial gracias a su capacidad para capturar relaciones contextuales entre palabras mediante mecanismos de atención, mejorando significativamente

el rendimiento en tareas como clasificación de texto, análisis de sentimientos o detección de sesgo (Vaswani et al., 2017).

Más concretamente, utilizaremos el modelo *bias-detection-model*, disponible en la plataforma Hugging Face [44], que ha sido entrenado para detecar sesgo en textos en inglés. Este modelo devuelve una puntuación numérica entre 0 y 1, donde valores más altos indican una mayor probabilidad de que la consulta sea considerada sesgada (véase Figura 39).

Figura 39. Ejemplo de código Python para el cálculo de sesgo.

```python
#MODELO bias-detection-model
tokenizer = AutoTokenizer.from_pretrained("d4data/bias-detection-model")
model = TFAutoModelForSequenceClassification.from_pretrained("d4data/bias-detection-model")
classifier = pipeline('text-classification', model=model, tokenizer=tokenizer) # cuda = 0,1 based on
gpu availability

classifier("The irony, of course, is that the exhibit that invites people to throw trash at vacuuming
Ivanka Trump lookalike reflects every stereotype feminists claim to stand against, oversexualizing
Ivanka's body and ignoring her hard work.")

#SALIDA

[{'label': 'Biased', 'score': 0.9819144606590271}]
```

Python

Fuente: Elaboración propia.

Figura 40. Primeras filas de DataFrame con etiquetas y puntuación de sesgo (2020).

	question	category	bias_label	bias_score
0	Can vitamin D cure COVID-19?	harmful	Non-biased	0.593567
1	Can vitamin D cure COVID-19?	harmful	Non-biased	0.593567
2	Can vitamin D cure COVID-19?	harmful	Non-biased	0.593567
3	Can vitamin D cure COVID-19?	harmful	Non-biased	0.593567
4	Can vitamin D cure COVID-19?	harmful	Non-biased	0.593567
...	...	...	...	...

Fuente: Elaboración propia.

[44] Hugging Face es un repositorio ampliamente utilizado por la comunidad de procesamiento del lenguaje natural para acceder a modelos preentrenados de IA y herramientas open source.

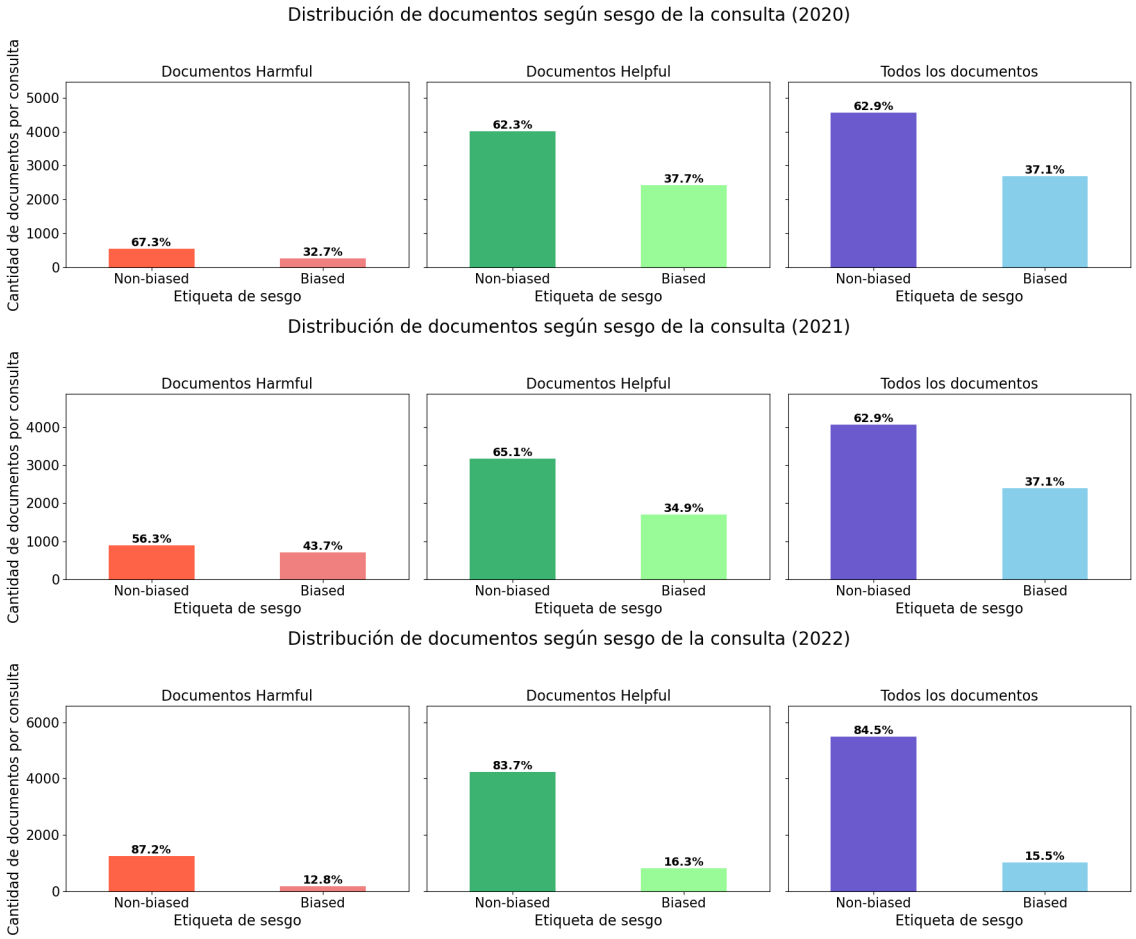
3.2.2 Resultados obtenidos del análisis de sesgo

3.2.2.1 Distribución de sesgo por año

En los tres años analizados, observamos que las consultas que recuperan mayor número de documentos útiles tienden a estar formuladas con un lenguaje menos sesgado.

Aunque en términos absolutos hay más documentos útiles, debemos fijarnos en la proporción dentro de cada categoría. Para el año 2020 y 2021, la proporción de documentos útiles no sesgados es superior a la proporción de documentos perjudiciales con la misma etiqueta. En cambio, en la edición de 2022, el panorama cambia ligeramente: la diferencia entre categorías se reduce y los documentos perjudiciales parecen ir perdiendo esa asociación con consultas sesgadas (véase Figura 41).

Figura 41. Distribución de documentos *harmful/helpful* según el sesgo de las consultas (2020–2022).



Fuente: Elaboración propia.

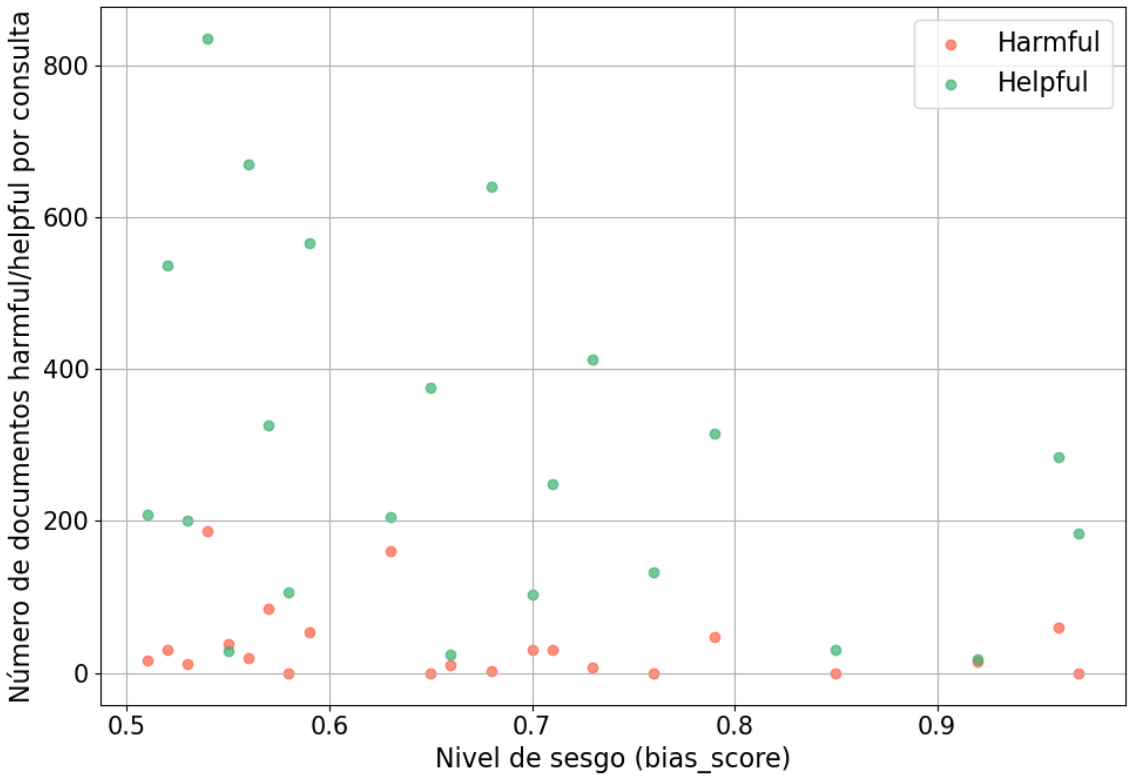
3.2.2.2 Análisis de dispersión

En general, los resultados obtenidos en los tres años analizados indican una relación débil entre el nivel de sesgo y el tipo de documentos recuperados, ya que las gráficas de dispersión muestran una alta dispersión de puntos y no permiten identificar patrones evidentes.

En la edición de 2020, se aprecia una ligera tendencia negativa con respecto a los documentos útiles, ya que tienden a aumentar de volumen conforme bajan los niveles de sesgo. Esto sugiere que las consultas con menor sesgo recuperan más contenido útil y menos perjudicial. Además, podemos ver como los puntos rojos correspondientes a documentos perjudiciales se concentran en zonas bajas del eje vertical, lo que refleja también que su volumen es mucho menor que el de los documentos útiles (véase Figura 42).

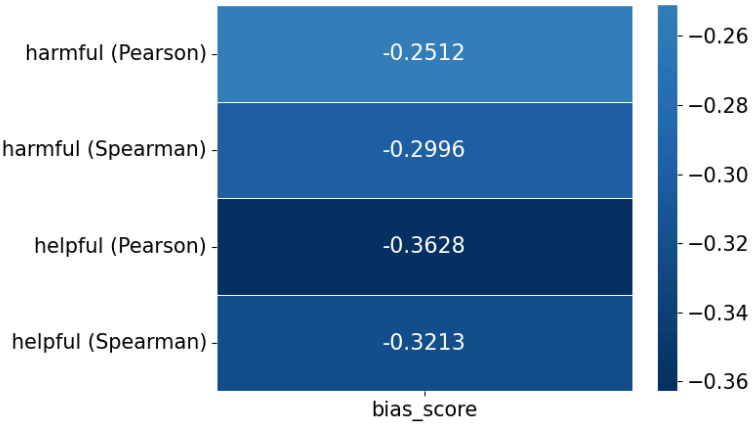
Con respecto a las correlaciones calculadas (Figura 43), podemos observar que todas las correlaciones entre *bias_score* y la cantidad de documentos (tanto Pearson como Spearman) son negativas. La correlación es más fuerte para documentos útiles, especialmente en Pearson (- 0.36), lo que nos indica una relación de tipo lineal.

Figura 42. Nivel de sesgo y número de documentos *helpful/harmful* por consulta (2020).



Fuente: Elaboración propia.

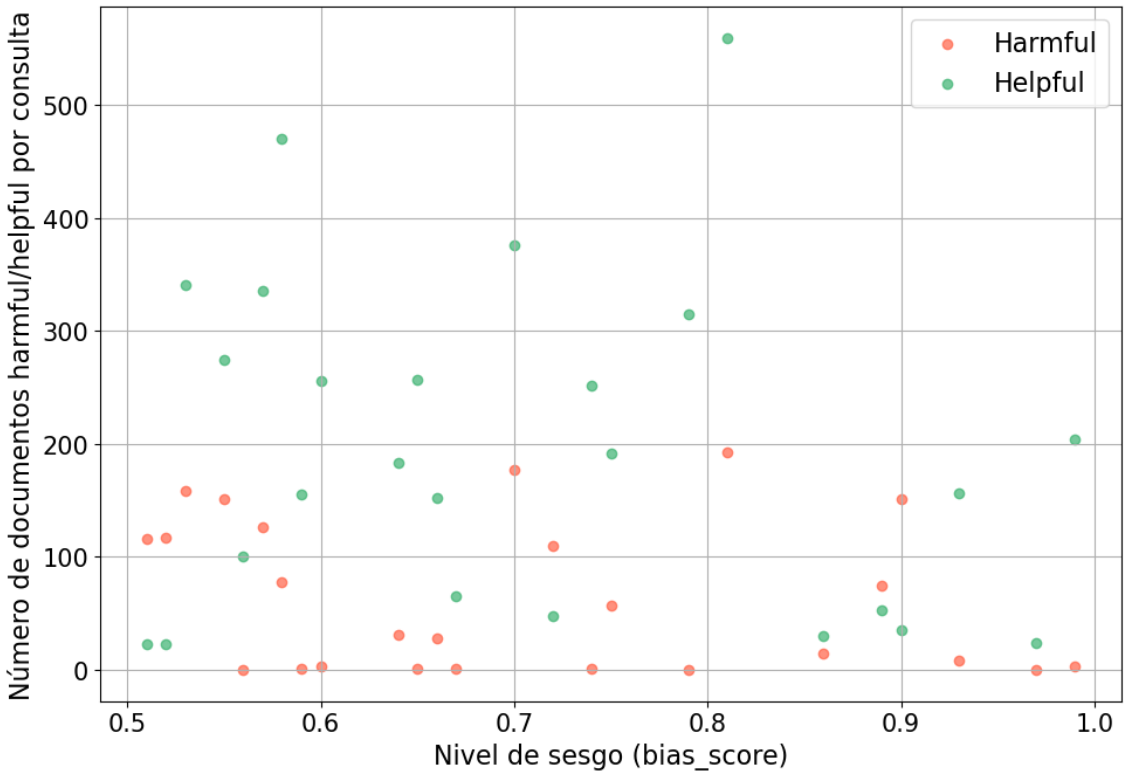
Figura 43: Correlaciones entre sesgo y cantidad de documentos *helpful/harmful* por consulta (2020).



Fuente: Elaboración propia.

En 2021, el patrón se mantiene, aunque de forma más débil. En la gráfica de dispersión (véase Figura 44) vemos como se mantiene la tendencia general en la que los documentos útiles se concentran más en los niveles bajos del eje de sesgo y alcanzan mayores volúmenes, mientras que los documentos perjudiciales están más distribuidos y rara vez superan valores altos en número. Aun así, la separación entre categorías es más difusa, y no se aprecia una tendencia visual tan clara como en el año anterior.

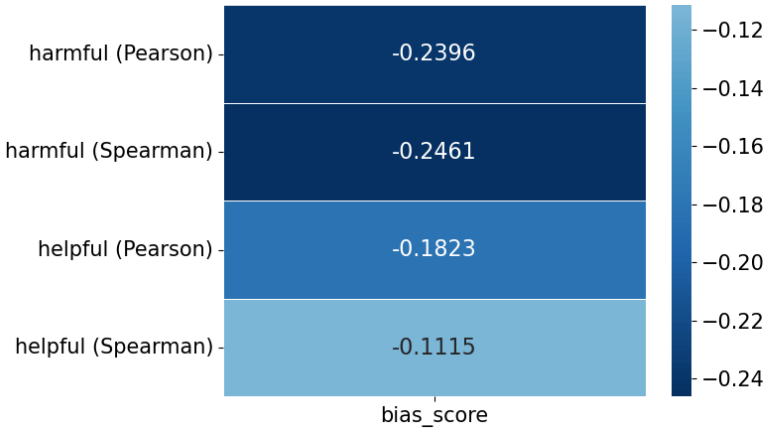
Figura 44: Nivel de sesgo y número de documentos *helpful/harmful* por consulta (2021).



Fuente: Elaboración propia.

Las correlaciones calculadas para la edición de 2021 confirman esta pérdida de señal para documentos útiles (véase Figura 45). De hecho, en este caso se sitúa por debajo de las correlaciones de documentos perjudiciales, que se mantiene en niveles similares al año anterior.

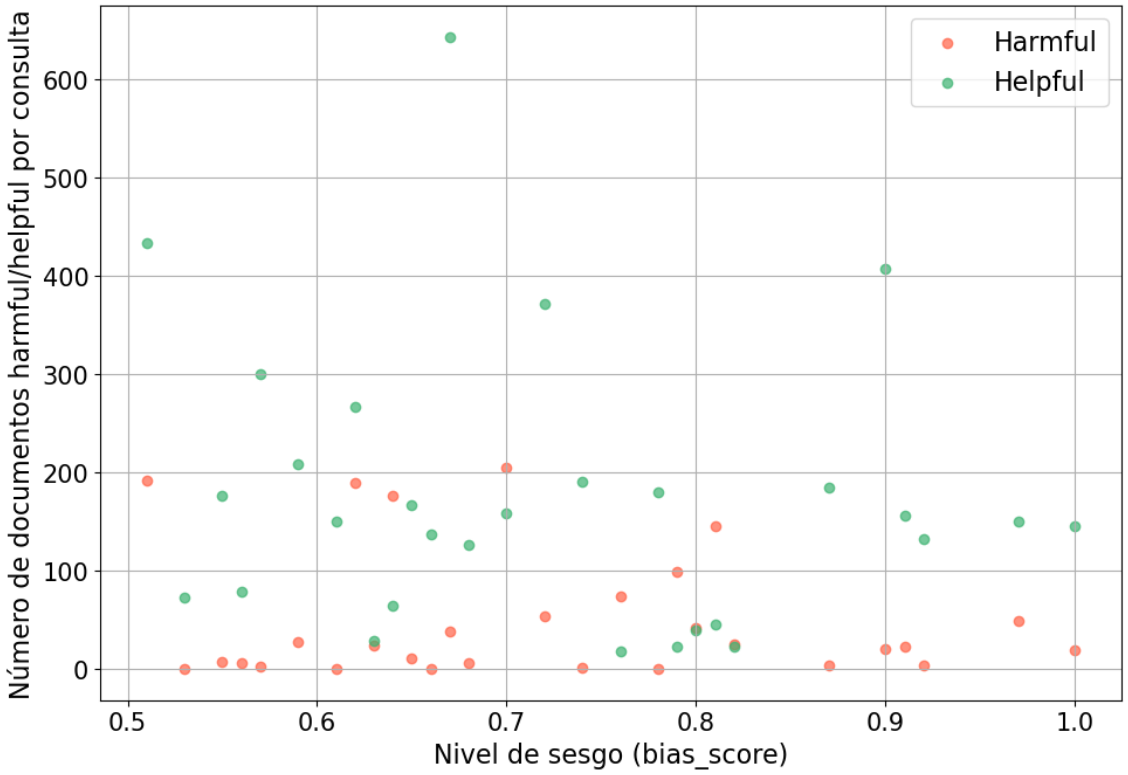
Figura 45. Correlaciones entre sesgo y cantidad de documentos *helpful/harmful* por consulta (2021).



Fuente: Elaboración propia.

En la edición de 2022, el patrón observado en años anterior pierde todavía más definición. Aunque en la gráfica de dispersión (Figura 46) seguimos viendo que los documentos útiles suelen alcanzar valores más altos en volumen y que los documentos perjudiciales se mantienen en niveles bajos del eje vertical, la separación entre ambas categorías se vuelve más difusa.

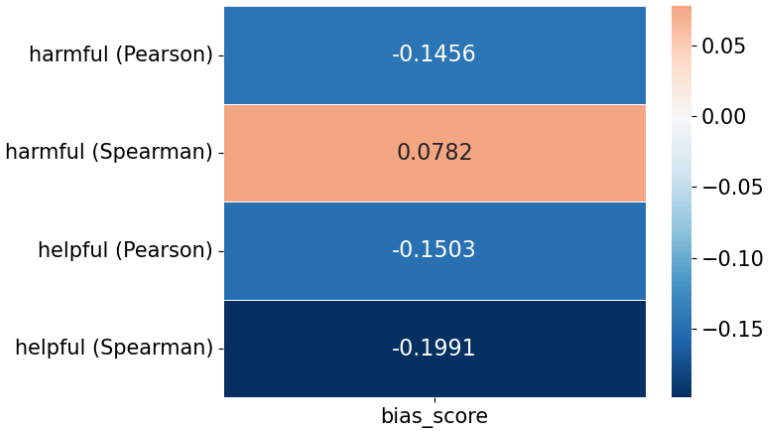
Figura 46. Nivel de sesgo y número de documentos *helpful/harmful* por consulta (2022).



Fuente: Elaboración propia.

En cuanto a las correlaciones (véase Figura 47), los valores son mucho más bajos y cercanos a cero, especialmente en el caso de los documentos perjudiciales. En este año incluso aparece una correlación Spearman positiva en *harmful* (+0.0782), lo que rompe con el patrón visto en años anteriores. Las correlaciones más fuertes se mantienen en la categoría *helpful*, con valores negativos tanto en Pearson como en Spearman, muy similares a los valores obtenidos en la edición de 2021.

Figura 47. Correlaciones entre sesgo y cantidad de documentos *helpful/harmful* por consulta (2022).



Fuente: Elaboración propia.

3.2.2.3 Conclusiones

Los resultados obtenidos a lo largo de los tres años reflejan que, aunque existe cierta relación entre el sesgo de una consulta y la calidad de los documentos recuperados, esta relación es débil e inconsistente.

En 2020, los datos sugerían que un nivel bajo de sesgo estaba asociado con un mayor volumen de documentos útiles, y las correlaciones obtenidas respaldaban esta idea, especialmente en el caso de los documentos útiles. Sin embargo, en los años siguientes la señal fue perdiendo fuerza.

Comparado con el análisis de frecuencia léxica, el modelo de sesgo ofrece un rendimiento inferior. Mientras que la frecuencia media y mínima permitieron detectar señales, sobre todo en el año 2021, el clasificador de sesgo apenas logra capturar patrones útiles en ningún año.

Esto podría deberse a los datos con los que fue entrenado el clasificador. Aunque no se disponen de todos los detalles técnicos, es probable que el modelo haya sido entrenado con textos del ámbito político, ya que es un dominio común para tareas de detección de sesgo. En nuestro caso, en cambio, se trabaja con consultas cortas y del ámbito médico, lo que puede hacer que el modelo no capture bien los matices propios de este tipo de lenguaje.

3.3 Modelos Grandes de Lenguaje

3.3.1 Introducción y modelo utilizado

Para terminar, se explora un enfoque diferente para predecir si una consulta médica puede estar relacionada con la recuperación de contenido útil o perjudicial. Para los anteriores predictores, nos basamos en características del lenguaje como la frecuencia de las palabras o el nivel de sesgo, pero en este caso se recurre a un Modelo Grande de Lenguaje (MGL). El objetivo es utilizar un modelo ya entrenado con gran conocimiento sobre el lenguaje, para probar si, simplemente leyendo la consulta, es capaz de estimar su riesgo de generar desinformación.

El modelo utilizado es *Mistral* de la librería de Python, Ollama. A cada consulta se le aplica un *prompt* para pedir directamente el grado de controversia o posible desinformación que podría generar cada consulta dada. A partir de la respuesta generada por el modelo, se

extrae una puntuación numérica que servirá de indicador de controversia, y nos será útil para los siguientes análisis.

3.3.2 Elección del *prompt* y aplicación del modelo

Para aplicar el modelo de lenguaje, primero es necesario definir de forma precisa la tarea que queremos que realice. En este caso, buscamos que el modelo evalúe hasta qué punto una consulta médica puede resultar problemática en términos de Recuperación de Información.

Tras probar varios *prompts,* se optó por una versión explícita que define claramente qué se entiende por controversia, proporcionando al modelo una escala del 1 al 5 (véase Figura 48).

Figura 48. *Prompt* utilizado para estimar el grado de controversia de las consultas médicas.

```
prompt3 = """You are an expert in health-related search engine results and misinformation detection.

Given a health-related query, assess its risk of returning harmful results (containing misinformation,
pseudoscience, or misleading health claims) and/or suppressing helpful results (accurate, evidence-based medical
information).

Controversial queries often trigger polarized opinions, recurring myths, or conflicting guidance from reliable
sources. Rate the level of controversy on a scale from 1 to 5 based on the following:

1 = Clear, factual query with consistent evidence and no common misinformation
2 = Mostly factual with minimal risk of misinformation or confusion
3 = Some disagreement or presence of minor myths or fringe opinions
4 = High presence of misinformation, conflicting evidence, or misleading narratives
5 = Very high controversy with widespread misinformation or dangerous claims

Just reply with a single number (1-5). Do not explain your answer.

Query: {query}"""
✓ 0.0s                                                                                    Python
```

Fuente: Elaboración propia.

A partir de ese *prompt,* se aplicó el modelo a todas las consultas de los conjuntos de datos de 2020, 2021 y 2022. El resultado fue una nueva columna con el grado de controversia obtenido con el modelo tal como se ve en la Figura 49.

Figura 49. Algunas consultas con puntuación de controversia y cantidad de documentos (2020)

question	harmful	helpful	controversy_score
Can 5G antennas cause COVID-19?	35	123	5
Can ACE and ARBs worsen COVID-19?	6	167	4
Can BCG vaccine prevent COVID-19?	0	151	4
Can Cannabis help COVID-19?	0	90	4
Can Echinacea prevent COVID-19?	53	56	4
Can Ginger cure COVID-19?	30	103	4
Can Hib vaccine prevent COVID-19?	3	149	3
Can Homemade Vodka Sanitizer prevent COVID-19?	13	273	5
Can Hydroxychloroquine worsen COVID-19?	0	183	4
Can IVIG cure COVID-19?	7	6	4

Fuente: Elaboración propia.

3.3.3 Resultados obtenidos

3.3.3.1 Análisis de distribución por niveles de controversia

Los datos de la edición de 2020, representados en diagramas de caja[45], muestran que a mayor nivel de controversia tiende a reducirse el volumen de documentos útiles y aumentar el de perjudiciales (Figura 50). Ese patrón es más evidente en el caso de los documentos útiles, donde las consultas con niveles bajos de controversia presentan las medianas[46] más altas y a medida que el nivel de controversia aumenta, la mediana desciende.

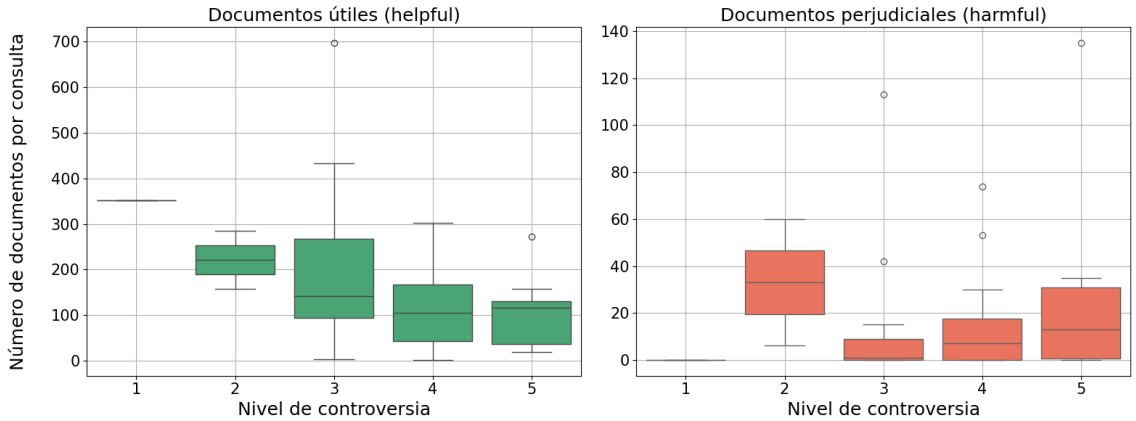
Se observa también que, en el nivel de controversia 3, el diagrama de caja presenta el rango intercuartílico[47] más amplio de todos los niveles. Esto significa que las consultas dentro de este nivel de controversia varían mucho entre sí en cuanto al número de documentos recuperados, algunas tienen un volumen bajo, y otras un volumen muy alto de documentos.

[45] A diferencia de los análisis realizados para los anteriores predictores en los que hemos utilizado gráficas de dispersión para explorar la relación entre el grado de controversia estimado por el modelo y la cantidad de documentos recuperados de cada tipo (*helpful* o *harmful),* en este caso se toma la decisión de usar diagramas de caja (*boxplots),* debido a que el nivel de controversia estimado es una variable categórica (del 1 al 5) y este tipo de visualización resulta más adecuado para representar con claridad la distribución de valores para cada categoría.

[46] La mediana representa el valor central del número de documentos recuperados. Es decir, divide las consultas en dos mitades: el 50 % tiene un número de documentos igual o inferior a la mediana, y el otro 50% igual o superior.

[47] El rango intercuartílico mide la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3), es decir, abarca el 50 % de las consultas con volúmenes de documentos situados en la parte media de la distribución.

Figura 50. Diagramas de caja de documentos útiles (verde) y perjudiciales (rojo) por nivel de controversia (2020).



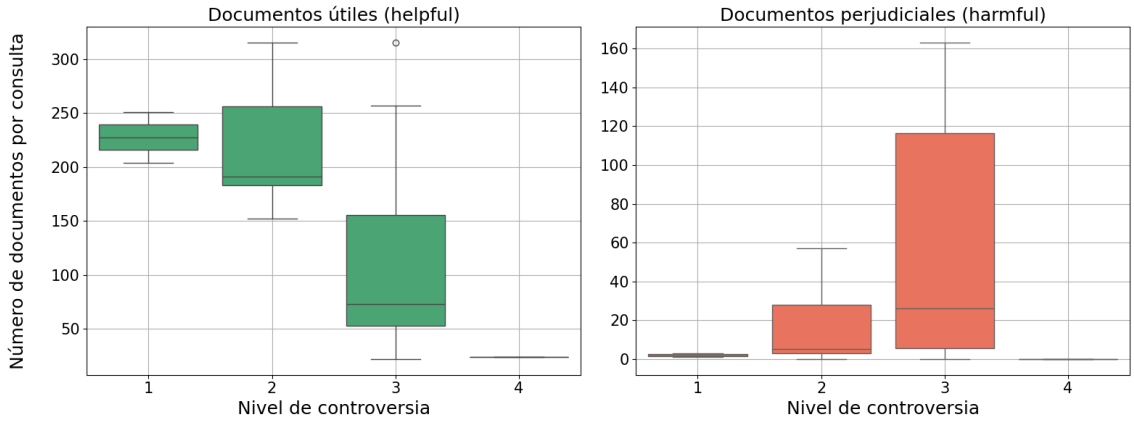
Fuente: Elaboración propia.

Por otro lado, los documentos perjudiciales muestran un patrón diferente. A simple vista sí se aprecia una relación positiva entre el nivel de controversia y el volumen de documentos de tipo *harmful,* ya que conforme aumenta el nivel de controversia, la mediana tiende a aumentar. Sin embargo, este patrón se ve alterado por el nivel 2 de controversia, que presenta la mediana más alta de todo el conjunto (véase Figura 50).

En los diagramas se observa también que, para esta edición hay una única consulta que el modelo ha etiquetado con el nivel de controversia 1. Se trata de *"Can handwashing prevent COVID-19?"*, que no tiene documentos perjudiciales asociados y presenta uno de los valores más altos en volumen de documentos útiles (352). Este caso refuerza la coherencia del modelo, al identificarla como una consulta con riesgo de desinformación prácticamente nulo y claramente vinculada a información útil.

En la edición de 2021 (véase Figura 51) y en la edición de 2022 (véase Figura 52), los datos mantienen en gran medida la tendencia observada en 2020 tanto para documentos útiles como perjudiciales. A diferencia de lo que ocurría en 2020, donde solo había una consulta en nivel de controversia 1, en estas dos ediciones si encontramos varias en documentos útiles. A medida que el nivel de controversia sube, la mediana de volumen de documentos útiles va descendiendo de una forma muy clara, y en el caso de los documentos perjudiciales, sucede lo contrario.

Figura 51. Diagramas de caja de documentos útiles (verde) y perjudiciales (rojo) por nivel de controversia (2021).

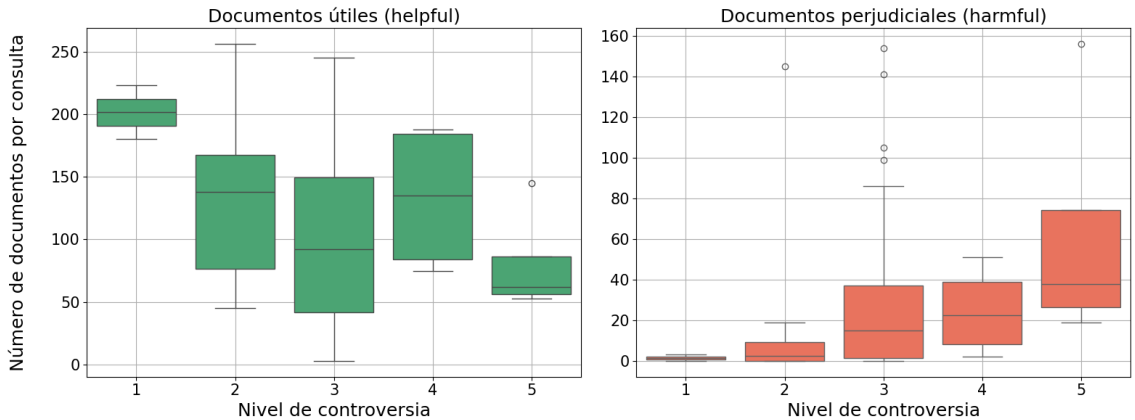


Fuente: Elaboración propia.

En la edición de 2021, se observa que no hay consultas etiquetadas con nivel 5 de controversia y solo hay una consulta en nivel 4. Además, tanto en documentos útiles como los documentos perjudiciales, el nivel 3 es el que muestra mayor dispersión y el rango intercuartílico más amplio.

En cambio, en 2022 si hay datos en todos los niveles de controversia, desde el 1 hasta el 5. Sin embargo, en esta edición encontramos un nivel que rompe con la tendencia negativa de los documentos útiles. Concretamente, el nivel 4 de controversia presenta una mediana de documentos útiles más alta que el nivel 3 y muy cercana a la altura del nivel 2.

Figura 52. Diagramas de caja de documentos útiles (verde) y perjudiciales (rojo) por nivel de controversia (2022).



Fuente: Elaboración propia.

Al analizar este caso concreto, se comprueba que esto se debe a la presencia de varias consultas que, a pesar de estar etiquetadas con nivel 4 de controversia, tienen asociado un

mayor volumen elevado de documentos útiles (véase Figura 53). Esto refleja que, aunque estas consultas presentan cierto grado de controversia, no implica necesariamente que recuperarán más documentos perjudiciales.

Figura 53. Consultas con nivel 4 de controversia y su distribución de documentos (2022).

```
display(qrels_2022_prompt3[qrels_2022_prompt3["controversy_score"] == 4])
```

✓ 0.0s Python

category	question	harmful	helpful	controversy_score
5	Are vaccines linked to autism?	51	183	4
20	Can grapefruit interfere with medication?	2	188	4
23	Can statins cause permanent cognitive impairment?	35	75	4
25	Can vape pens be harmful?	10	87	4

Fuente: Elaboración propia.

3.3.3.2 Correlaciones

Las correlaciones obtenidas entre el nivel de controversia y el volumen de documentos útiles y perjudiciales refuerzan las tendencias observadas en los diagramas de caja.

En el caso de los documentos útiles, se observa de forma consistente una correlación negativa en los tres años (véase Figura 54). Esta relación es especialmente fuerte en la edición de 2021, donde la correlación de Pearson alcanza -0.60 y la de Spearman llega hasta -0.65, marcando una tendencia más monótona que lineal (Spearman con valor más alto que Pearson).

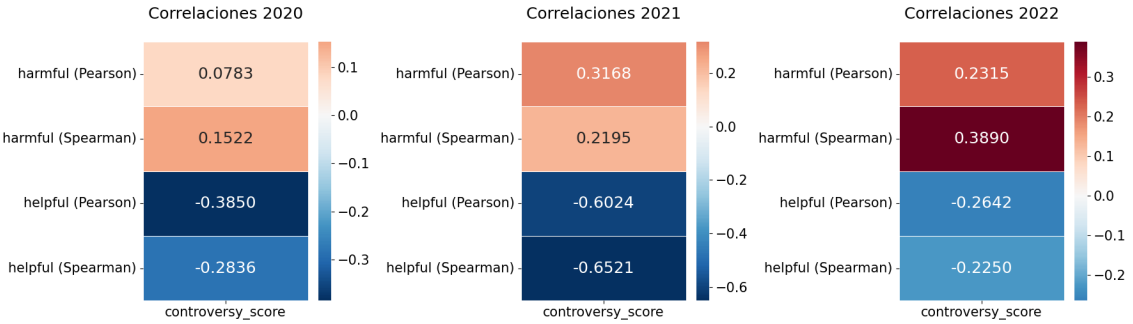
En las ediciones de 2020 y 2022, aunque las correlaciones son más moderadas (-0.38 y -0.26 en Pearson), mantienen igualmente esta relación inversa.

Con respecto a los documentos perjudiciales, las correlaciones son todas positivas, lo que confirma esa tendencia a aumentar el volumen de desinformación a medida que aumenta la controversia. Esta relación es más lineal en el año 2021 (valor en Pearson más alto que Spearman) y más monótona en 2020 y 2022.

En la edición de 2020, donde las correlaciones para documentos perjudiciales son mucho más bajas. Esto puede ser debido a la temática específica de ese año relacionada con el COVID-19 que, al igual que en los análisis de los predictores anteriores, ha afectado los resultados de forma negativa. Además, como ya analizamos en los diagramas de ese año (Figura 50), hay comportamiento diferente de los documentos etiquetados en el nivel 2 que hace que rompa con la tendencia positiva vista en las otras dos ediciones, y eso puede influir en los

resultados. Lo mismo ocurre con las correlaciones en documentos útiles de la edición de 2022 que bajan en comparación con el 2021, también influido por esa rotura en el patrón observado en el nivel 4 de controversia (Figura 52).

Figura 54. Correlaciones entre nivel de controversia y documentos *helpful/harmful* por consulta.



Fuente: Elaboración propia.

3.3.3.3 Conclusiones

En general, los resultados obtenidos por el predictor basado en modelos de lenguaje son buenos y coherentes. A diferencia de lo que ocurría con los otros predictores analizados, con este encontramos un patrón muy claro que se repite en los tres años: las correlaciones son siempre negativas en documentos útiles y positivas en documentos perjudiciales. Esta coherencia es, sin duda, uno de los puntos fuertes de este predictor.

Además, las correlaciones que hemos visto encajan bastante bien con lo visto en los diagramas de caja en los que a medida que sube la controversia, baja la mediana de documentos útiles y sube la de perjudiciales.

Conclusiones y ampliación

En este trabajo se ha analizado la posibilidad de predecir el riesgo de recuperación de desinformación en los resultados de búsqueda médica a partir de las características de la propia consulta. Para ello, se han evaluado tres tipos de predictores aplicados a los datos del TREC *Health Misinformation Track* (2020-2022).

El análisis de frecuencia léxica mostró ciertas señales, especialmente en 2021, donde las consultas formuladas con palabras más comunes tienden a recuperar más documentos perjudiciales, mientras que las consultas con términos técnicos recuperan mayor proporción de contenido útil. Sin embargo, este patrón no fue constante en todos los años, lo que limita su robustez como predictor único.

El modelo de detección de sesgo ofreció un rendimiento inferior. Aunque se detecta cierta relación entre consultas menos sesgadas y mayor volumen de documentos útiles, la señal es débil y no consistente. Probablemente, el bajo rendimiento se debe a que este tipo de modelos no está optimizado para consultas breves ni para el contexto médico.

Por el contrario, el predictor basado en Modelos Grandes de Lenguaje (MGL) demostró ser el más fiable y coherente. En los tres años analizados se mantiene un patrón claro: a mayor nivel de controversia en la consulta, aumenta la cantidad de documentos perjudiciales y disminuye la de útiles. Este comportamiento se refleja tanto en los diagramas de distribución como en las correlaciones obtenidas.

Los resultados de este estudio podrían aplicarse en el desarrollo de herramientas que ayuden a combatir la desinformación en buscadores y plataformas online. Una aplicación práctica sería implementar un sistema de alerta capaz de detectar si la consulta realizada por un/a usuario/a presenta un riesgo elevado de atraer información perjudicial.

Como propuesta, se podría desarrollar una **extensión de navegador** que, al detectar que un/a usuario/a introduce una consulta médica, analice automáticamente su grado de controversia y ofrezca un aviso sobre el riesgo de desinformación. Esta herramienta permitiría a los/las usuarios/as ser más conscientes de los posibles riesgos asociados a sus búsquedas y favorecería una navegación más segura, especialmente en temas de salud.

Bibliografía

BBC News Mundo. (2020, 29 de mayo). Coronavirus y desinformación: por qué algunos remedios falsos y teorías conspirativas pueden ser peligrosos para tu salud. *BBC News Mundo.* Recuperado el 9 de junio de 2025, de https://www.bbc.com/mundo/noticias-52840201

Clarke, C. L. A., Maistro, M., Rizvi, S., Smucker, M. D., e Zuccon, G. (2020). *Overview of the TREC 2020 Health Misinformation Track.*

Clarke, C. L. A., Maistro, M., y Smucker, M. D. (2021). *Overview of the TREC 2021 Health Misinformation Track.*

Clarke, C. L. A., Maistro, M., Seifikar, M., y Smucker, M. D. (2022). *Overview of the TREC 2022 Health Misinformation Track.*

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine*, 113, 763–765.

Fox, S. (2011). *80% of internet users look for health information online*. Pew Internet & American Life Project. Recuperado el 6 de junio de 2025, de https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/2003/PIP_Health_Report_July_2003.pdf.pdf

National Center for Health Statistics. (2023). *Use of the internet for health information among adults: United States, 2022* (Data Brief nº 482). Recuperado el 8 de junio de 2025, de https://www.cdc.gov/nchs/products/databriefs/db482.htm

Pogacar, A., Ghenai, A., Smucker, M. D., e Clarke, C. L. (2017). The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 209–216).

Reuters Institute. (2021). *Reuters Digital News Report 2021*. University of Oxford. Recuperado el 8 de junio de 2025, de https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021

Reuters Institute. (2024). *Digital News Report 2024 – Resumen ejecutivo*. Recuperado el 8 de junio de 2025, de https://reutersinstitute.politics.ox.ac.uk/es/digital-news-report/2024/dnr-resumen-ejecutivo

RTVE. (2024, 18 de octubre). *Día Mundial contra el Cáncer de Mama: Los bulos que ponen en riesgo la salud de las pacientes*. Recuperado el 8 de junio de 2025, de https://www.rtve.es/noticias/20241018/dia-mundial-contra-cancer-mama-bulos-ponen-riesgo-salud-pacientes/16290388.shtml

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need. In Advances in Neural Information Processing Systems* (Vol. 30).

Vigdor, N. (2020, 24 de marzo). Man fatally poisons himself while self-medicating for coronavirus, doctor says. *The New York Times*. Recuperado el 8 de junio de 2025, de https://www.nytimes.com/2020/03/24/us/chloroquine-poisoning-coronavirus.html

Weber Shandwick. (2023). *The Great American Search for Healthcare Information*. Recuperado el 8 de junio de 2025, de https://webershandwick.com/news/the-great-american-search-for-healthcare-information