La Dificultad de la Etiquetación de Desinformación: Un Caso de Estudio para Búsquedas Relacionadas con el Gas Radón

The Difficulty of Misinformation Labelling: A Case Study for Radon Gas-Related Searches

Noel Pascual-Presa[1], Marcos Fernández-Pichel[2], David Enrique Losada[2], Berta García-Orosa[1], Paula Martínez-Graña[1], Lucía Ortigueira-Piñeiro[1]

[1]*Departamento de Ciencias da Comunicación, Universidad de Santiago de Compostela, 15782, Santiago de Compostela, España*
[2]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, España*

Resumen

La creación de colecciones etiquetadas relacionadas con la desinformación es un aspecto crucial para impulsar el desarrollo de tecnologías automáticas que filtren contenidos nocivos. Esto es particularmente importante en riesgos relacionados con la salud. Sin embargo, la asignación de etiquetas de calidad (por ejemplo, correctitud o credibilidad) a los textos es algo que debe realizarse de manera rigurosa. En este artículo describimos nuestros esfuerzos para crear una colección de pasajes etiquetados en referencia a su relevancia y calidad para búsquedas relacionadas con los riegos para la salud del gas radón. Además de ilustrar las dificultades encontradas en un proyecto de etiquetación de esta índole, con este trabajo contribuimos mediante la puesta a disposición de la comunidad científica de un nuevo recurso anotado, que puede ser explotado en el futuro para impulsar aprendizaje supervisado en este ámbito.

Palabras clave

Búsqueda Web, Desinformación, Gas Radón, Etiquetación

Abstract

The creation of labelled collections related to misinformation is a crucial aspect in the development of automatic technologies that filter harmful content. This is particularly important for health-related risks. However, assigning quality labels (e.g., correctness or credibility) to texts needs to be done rigorously. In this article, we describe our endeavours to build a collection of labelled passages, with relevance and quality annotations, for search tasks related to the risks of radon gas. In addition to illustrating the difficulties encountered in a labelling project of this kind, our contribution with this work is to provide the scientific community with a new annotated resource that can be used in the future to support supervised learning in this area.

Keywords 1

[1] [1]SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

EMAIL: noel.pascual.presa@usc.es (N.Pascual-Presa); marcosfernandez.pichel@usc.es (M.Fernández-Pichel); david.losada@usc.es (D.E.Losada); berta.garcia@usc.es (B.García-Orosa); paula.martinez.grana@gmail.com (P.Martínez-Graña) lucia.ortigueira@rai.usc.es (L.Ortigueira-Piñeiro)

ORCID: 0009-0002-9091-7631 (N.Pascual-Presa); 0000-0002-6560-9832 (M. Fernández-Pichel); 0000-0001-8823-7501 (D.E. Losada) 0000-0001-6126-7401 (B.García-Orosa); 0000-0003-0769-4159 (P.Martínez-Graña); 0009-0007-6415-6337 (L.Ortigueira-Piñeiro)

Web Search, Misinformation, Radon Gas, Labelling.

1. Introducción

Internet se ha convertido en un valioso recurso que ha reconfigurado la manera en la que las personas accedemos y comprendemos la información [1]. Hoy en día los recursos en línea son la fuente principal para la consulta de información relacionada con la salud [2]. Esto otorga a los buscadores web una gran responsabilidad y peso en el acceso a información sobre este tema [3]. Por ejemplo, los resultados de una búsqueda relacionada con la salud pueden motivar cambios en los comportamientos de las personas frente a una determinada amenaza sanitaria [4]. A su vez, este acceso a grandes volúmenes de datos que está al alcance de una gran parte de la población mundial trae consigo una serie de riesgos [5]. La cantidad de información falsa y desinformación aumenta constantemente [6] y esto supone una grave amenaza y peligro para la población, especialmente, cuando se trata de información relacionada con la salud. La información falsa en línea no solo afecta a la confianza que deposita el usuario en las evidencias científicas, sino que también puede motivar decisiones de salud perjudiciales o contraproducentes para los ciudadanos [7]. Es por ello por lo que es necesario realizar nuevos esfuerzos de investigación que busquen construir entornos en línea de confianza y seguros, disminuyendo la cantidad de información falsa o desinformación [8].

Uno de los principales problemas a los que nos enfrentamos en este campo es la gran cantidad de datos disponibles y la dificultad de distinguir la información veraz de la desinformación [9]. En este aspecto, los algoritmos de recuperación y clasificación de información son una herramienta muy útil para tratar de contrarrestar este problema. Estos algoritmos son capaces de recuperar y detectar desinformación, pero, por lo general, ven afectado su rendimiento y precisión por calidad de los datos de entrenamiento y prueba proporcionados [10]. Generar conjuntos de datos valiosos y de calidad para entrenar estos modelos, suele ser un proceso costoso y complejo, que habitualmente requiere la participación de asesores humanos y unos criterios de clasificación minuciosos y detallados. Esta investigación nace con el objetivo de generar un recurso sólido, transparente y robusto con miles de páginas webs que pueda ser utilizado para otros estudios en diferentes ámbitos como las Ciencias de la Computación o la Comunicación. Además, se busca ofrecer nuevas pautas y criterios que permitan establecer un juicio sólido para el análisis de la calidad de la información relacionada con la salud.

Para ello, este estudio ha sido llevado a cabo por un equipo interdisciplinar, compuesto por investigadores del campo de la ingeniería informática, comunicación y periodismo, partiendo de la necesidad de incorporar distintas técnicas de investigación y así poder desarrollar un método que permita analizar grandes volúmenes de información relacionada con la salud en Internet. En esta ocasión, se ha optado por basar la investigación en un estudio de caso sobre la información relacionada con los riesgos para la salud del gas radón que, por las características de este riesgo, lo hace idóneo para esta ocasión.

La contribución de este estudio es, por tanto, dual:

- Por una parte, se definen y evalúan una serie de criterios para el etiquetado de desinformación web relacionada con la salud, en concreto, sobre el gas radón.
- Por otra parte, se crea un recurso etiquetado que puede ser utilizado por otros grupos de investigación. Se trata de un corpus de páginas web anotadas en términos de su relevancia y calidad que figuran como resultados de búsqueda para consultas relacionadas con los riesgos del gas radón. Por tanto, este valioso recurso de etiquetado lo ponemos a disposición de la comunidad científica previa solicitud a los autores en caso de que el artículo sea aceptado.

2. Trabajo relacionado

La creación de recursos etiquetados es la base para el desarrollo de algoritmos supervisados en el ámbito del Procesamiento del Lenguaje Natural (PLN). Normalmente, la creación de estos recursos implica la intervención de asesores humanos, lo que añade un componente de subjetividad que puede llevar a desacuerdos y a una etiquetación inadecuada [11]. Para paliar este efecto, la creación de recursos etiquetados suele basarse en la definición de una serie de guías o criterios que los

etiquetadores deben seguir. Esta es la principal técnica de creación de conjuntos "golden truth" utilizada en campos como el PLN o la Recuperación de la Información (RI) e implementada regularmente en prestigiosas conferencias que liberan recursos textuales como la Text Retrieval Conference (TREC) o la propia conferencia de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

Sin embargo, estudios previos han demostrado que algunos procesos de anotación carecen de la rigurosidad necesaria para realizar una asignación objetiva de las etiquetas. Por ejemplo, algunas investigaciones previas centraron sus esfuerzos en definir una serie de guías capaces de generar un etiquetado robusto y objetivo [12]. En Fernández-Pichel et al [10], los autores demostraron la inconsistencia de los juicios de credibilidad generados siguiendo las recomendaciones oficiales de la tarea "Health Misinformation Track" dentro de TREC. Como alternativa, propusieron una serie de guías que mejoraban el acuerdo entre asesores y producían un etiquetado robusto de la dimensión de la credibilidad de la información. Otros autores [13] también propusieron una serie de guías para la generación de recursos etiquetados destinados al entrenamiento de algoritmos para la detección de desinformación médica. Por otra parte, en otras investigaciones [14] se definieron una serie de indicadores de la credibilidad de artículos, que incluían indicadores complejos como falacias lógicas o el tono del discurso.

En este estudio, la contribución es dual: por una parte, se continúa con la línea abierta por investigaciones previas y se definen una serie de guías para el etiquetado de relevancia y de calidad de la información de un caso altamente especializado como el gas radón; por otra parte, se genera un recurso, a disposición de la comunidad, que puede servir a grupos investigadores de diferentes disciplinas.

3. Creación de una colección de pasajes web relacionados con riesgos sobre el gas radón

El radón es un gas noble radioactivo subproducto de la descomposición natural del uranio presente en suelo y rocas. La exposición a este gas está considerada como la primera causa de cáncer de pulmón en no fumadores y la segunda causa en fumadores [15]. La relevancia de este riesgo no solo radica en los graves efectos demostrados en la salud de las personas, sino también en su persistencia a lo largo del tiempo, que lo convierte en un riesgo atemporal. Esto nos permite un análisis ajeno a posibles crisis puntuales, como puede pasar con otros riesgos para la salud.

Una de las principales características de este gas radica en que está presente por todo el planeta. A pesar de ello, no todos los países se ven afectados de la misma forma por el radón, ya que sus niveles de presencia dependen mayoritariamente de la composición geológica del suelo. Otro factor que incrementa los riesgos del radón es su naturaleza insípida, incolora, e inodora, lo que lo convierte en indetectable para las personas a no ser que se realicen pruebas específicas [16]. Tanto la Organización Mundial de la Salud (OMS) como la Unión Europea han enfatizado en numerosas ocasiones la importancia de la verificación de la información sobre el radón y sus riesgos para la salud pública.

En esta ocasión, para la creación de una colección de pasajes web sobre el radón, en primer lugar, se han seleccionado 51 consultas relacionadas con los riesgos para la salud del radón. Estas consultas textuales han de ser representativas del tipo de consulta que un usuario convencional realizaría a un motor de búsqueda web. Para otorgar este realismo al recurso generado, se ha optado por utilizar búsquedas de información reales sobre los riesgos del radón. Para ello, todas las consultas seleccionadas se han obtenido de dos fuentes: i) los "query logs" de la *TREC Million Query Track* (2007, 2008 y 2009) y ii) diferentes cuestionarios realizados a la ciudadanía enmarcados en un proyecto financiado por el Consejo de Seguridad Nuclear de España sobre la percepción de la opinión pública del radón en España[2]. Además, la selección definitiva de las consultas a realizar fue llevada a cabo por un equipo de especialistas del área de la comunicación con experiencia previa en proyectos vinculados al gas radón, con el propósito de elaborar una muestra lo más representativa posible y que se ajustara a necesidades reales de búsqueda. Dada la limitada disponibilidad de consultas escritas en español, se ha optado por trabajar con necesidades de información escritas en inglés y todas aquellas

[2]Radón en España: percepción de la opinión pública, agenda mediática y comunicación del riesgo (RAPAC) del Consejo de Seguridad Nuclear (SUBV-13/2021)».

consultas fruto de los diferentes cuestionarios realizados a la ciudadanía se han traducido a ese idioma[3]. Algunos ejemplos de consultas son: *Radon causes cancer; How Radon Affects Children; How to reduce radon levels*.

Para simular los resultados de búsqueda obtenidos a partir de cada consulta seleccionada, se ha indexado un corpus masivo de páginas web, el denominado C4 [17] que está compuesto por millones de webs (1.590.000) indexadas en inglés y obtenidas de la web en abril de 2019[4]. Disponer de este corpus offline estático permite la replicación de este estudio y posibilita la comparación de algoritmos y variantes de búsqueda contra un repositorio centralizado [18, 19]. Para ejecutar estas búsquedas, se indexó el corpus en una estructura de índice invertido, análoga a la que utilizan los motores de búsqueda. Acto seguido, se buscaron páginas relevantes para las consultas previamente seleccionadas. Para ello, se emplearon técnicas de búsqueda estándar basadas en emparejamiento de palabras entre la consulta y los documentos. Concretamente, se utilizó el conocido algoritmo de búsqueda BM25 con su configuración por defecto en Pyserini [20]. Este modelo léxico ("sparse") puede no ser suficiente para encontrar los documentos más relevantes para una determinada búsqueda (debido a que, por ejemplo, no contempla sinonimia o similitud semántica o contextual). Por ello, se optó por reordenar el top 100 de documentos recuperados para cada consulta utilizando técnicas basadas en redes de neuronas profundas que estiman la similitud semántica entre la consulta y el documento [21]. Asimismo, para las 100 páginas más relevantes de cada consulta, realizamos extracción de pasajes relevantes dentro de esas páginas. Esto fue implementado con un algoritmo de IA de búsqueda de pasajes, que estima qué parte de la página es más central para responder a la consulta. Para esta parte, se utilizó el modelo MonoT5 entrenado para la detección de pasajes relevantes con la colección de MS MARCO (Microsoft Machine Reading Comprehension), una colección ampliamente utilizada a nivel mundial con pasajes etiquetados por relevancia [21].

El resultado de este proceso fue un conjunto de 5.100 pasajes relacionados con las consultas realizadas.

4. Criterios de evaluación de relevancia

Una vez generado el corpus a analizar, se procedió a la creación de unas guías de etiquetación que verifiquen la relevancia de cada uno de los pasajes recuperados. Esto se debe a que la búsqueda web en Internet es imperfecta, y podría darse el caso de que, por ejemplo, cierta información recuperada mediante estos procedimientos estándar de búsqueda y recuperación de la información no fuese relevante para la consulta. Por ello, una parte importante de esta investigación se centró en generar un **recurso consistente en pasajes webs etiquetados en cuanto a su relevancia** para cada una de las consultas seleccionadas previamente.

Para ello, se crearon previamente unos criterios para poder concretar si un pasaje debía ser considerado *muy relevante*, *parcialmente relevante* o *no relevante* para una determinada consulta relacionada con los riesgos del gas radón para la salud. Esta definición de unos criterios sólidos y robustos para la realización de juicios de relevancia consulta-documento es un método estándar en procesos de etiquetado de grandes volúmenes de datos en el campo de Recuperación de Información [10].

Para asegurar la calidad de los juicios de relevancia, se realizó un proceso de refinamiento en múltiples etapas para definir y mejorar las pautas de anotación. Inicialmente, se seleccionaron aleatoriamente 3 consultas (dentro del conjunto de 51 necesidades de información) y se analizaron los 100 pasajes recuperados para cada una de las consultas. Estos pasajes fueron procesados con una versión inicial de los criterios de relevancia por parte de tres evaluadores[5], que categorizaron cada pasaje en uno de los tres niveles de relevancia. Tras esta primera ronda, se calcularon métricas estándar para evaluar la concordancia entre los evaluadores. En concreto, se utilizó el Kappa de Cohen ponderado para evaluar la concordancia entre evaluadores individuales y el alpha de Krippendorff para evaluar la concordancia entre todos ellos. Los valores de Kappa oscilaron entre 0,46 y 0,62 con una mediana de 0,53. El alpha de Krippendorff arrojó un valor de 0,63. Tras examinar los comentarios de los evaluadores, identificamos discrepancias en la interpretación de las directrices entre los evaluadores. En concreto, los porcentajes

[3]Estas traducciones del español al inglés han sido llevadas a cabo por los autores de la investigación y revisadas por nativos.

[4]Para la indexación se utilizó la tecnología estándar de Pyserini: https://github.com/castorini/pyserini

[5]Los tres evaluadores, son investigadores del campo de la comunicación y la información vinculados a proyectos de I+D+I sobre el radón y la comunicación digital.

iniciales de acuerdo eran notablemente bajos debido a que no existía un criterio uniforme entre los etiquetadores para distinguir un pasaje como *parcialmente relevante* o como *muy relevante*. Por ello, se tuvo que llevar a cabo una reunión de grupo entre los asesores para unificar criterios en torno a juzgar un pasaje en función de su relevancia. Después de identificar estos motivos, se actualizaron las pautas y se repitió el proceso.

En una segunda iteración de etiquetación se logró un acuerdo más alto, siendo los valores de Kappa entre 0,71 y 0,73, y el alpha de Krippendorff de 0,83. Por tanto, se dio luz verde a estos criterios finales para llevar el etiquetado global de todos los pasajes de cada una de las consultas. El resultado es una "Guía de etiquetación de relevancia de información para consultas relacionadas con riesgos del radón para la salud".

- Irrelevante (0): el pasaje no responde a la consulta o necesidad de información. Por ejemplo, el pasaje habla de radón y/o de cáncer, pero no de una relación causal entre ambas.
- Parcialmente Relevante (1): el pasaje responde de manera parcial a la consulta o necesidad de información. Por ejemplo, el pasaje habla de la relación entre el radón y el cáncer y de la posibilidad de que el primero produzca al segundo, pero no da una información completa sobre el tema por el que pregunta el usuario.
- Muy relevante (2): el asesor encontrará la información del pasaje muy relevante si responde de manera muy clara a la necesidad de información. Por ejemplo, el pasaje contendrá una respuesta directa (incorrecta o no) de si el radón causa cáncer.

Es necesario tener en cuenta que en esta fase de la etiquetación no se evaluó ningún tipo de correctitud o calidad de la información, solo única y exclusivamente la relevancia de los pasajes para cada consulta.

El resultado final fue un recurso de 5.100 pasajes extraídos de la búsqueda web en Internet etiquetados en base a su relevancia para consultas relacionados con los riesgos del radón. Los datos obtenidos de este proceso de etiquetado muestran que el **56,68 %** de los pasajes anotados fueron considerados como *irrelevantes* para las necesidades de información de los usuarios. Este es un resultado normal, teniendo en cuenta que se analizaron las 100 primeras páginas webs recuperadas y la presencia de información "off-topic" es habitual en este tipo de búsquedas. Por otro lado, el **30,89 %** de los pasajes fueron percibidos como *parcialmente relevantes*, sin satisfacer de manera completa la necesidad de información, y el **14,33 %** de los pasajes presentaron información muy completa, es decir, se consideraron pasajes *muy relevantes*. Este recurso etiquetado de relevancia representa un hito significativo, con potencial para ser utilizado en diversas áreas. En concreto, esta colección resulta útil para una amplia gama de proyectos, como aquellos relacionados con el aprendizaje automático supervisado, y similar, por ejemplo, a MS MARCO.

5. Criterios de evaluación de calidad

En una segunda etapa de etiquetado, nos enfocamos a evaluar la calidad de los pasajes relevantes recuperados (limitándonos a aquellos considerados *parcialmente relevantes* o *muy relevantes* en la fase anterior). Para necesidades de información críticas como los riesgos del gas radón, es esencial considerar variables que ayuden a estimar la calidad de los pasajes extraídos. Por ejemplo, teniendo en consideración aspectos como la referencia a fuentes reputadas, la exactitud de la información proporcionada o la ausencia de contenido comercial. Tomando como referencia estudios pasados [10], se establecieron unos criterios iniciales basados en ciertos indicadores y, tras etiquetar los pasajes *parcialmente relevantes* o *muy relevantes* de tres consultas aleatorias, se calculó el acuerdo entre los evaluadores humanos. Como sucedió en la fase anterior de etiquetado, los valores de acuerdo iniciales eran demasiado bajos como para proceder con un etiquetado global. Tras una reunión de grupo para abordar discrepancias y reforzar las pautas, se repitió el proceso, logrando un mayor acuerdo y aumentando los valores de Kappa hasta 0,88-0,92 con una mediana de 0,9 y el alpha de Krippendorff hasta 0,90. Los criterios resultantes fueron consolidados para un etiquetado global de calidad de los pasajes. A partir de las anotaciones de los expertos, se definió un nivel de preferencia considerando el riesgo potencial para las personas. Por ejemplo, los contenidos más perjudiciales citan fuentes confiables, pero contienen información incorrecta y tienen propósitos comerciales. Estos pueden confundir a los usuarios y llevar a decisiones peligrosas. Por otro lado, los mejores contenidos son precisos, citan

fuentes confiables y carecen de intenciones comerciales. La siguiente tabla recoge los criterios de la "Guía de etiquetación de la calidad de información para consultas relacionadas con riesgos del radón para la salud". Esta guía ha sido desarrollada ad hoc por los autores de esta investigación, inspirándose en otras guías de preferencia documental presentes en la literatura [22].

Tabla 1
Niveles de calidad según los criterios "Guía de etiquetación de la calidad de información para consultas relacionadas con riesgos del radón para la salud"

Criterio	Cita fuente reputada	Intención comercial	Información correcta	Nivel calidad
1	✓	×	✓	3
2	✓	✓	✓	2
3	×	×	✓	2
4	×	✓	✓	1
5	×	×	×	-1
6	×	✓	×	-2
7	✓	×	×	-2
8	✓	✓	×	-3

- Criterio 1. El pasaje cita información proveniente de alguna de las siguientes procedencias: expertos/as reputados/as, artículos científicos, editoriales médicas, páginas de organismos gubernamentales, u otras fuentes similares que se estimen como autoritarias en la materia. Además, el pasaje no contiene información que contradice el consejo médico y tampoco contiene anuncios o intenciones de *marketing*. Ejemplo: "Según la OMS, el radón es uno de los principales causantes de cáncer de pulmón", "Según el NHS, una exposición prolongada al gas radón puede producir diversos problemas de salud".
- Criterio 2. El pasaje cita información proveniente de alguna de las siguientes procedencias: expertos/as reputados/as, artículos científicos, editoriales médicas, páginas de organismos gubernamentales, u otras fuentes similares que se estimen como autoritarias en la materia. Sin embargo, el pasaje contiene anuncios o intenciones de *marketing*. En todo caso, el pasaje no contiene información que contradice el conocimiento médico establecido. Ejemplo: "La OMS advierte que el radón es la segunda causa principal de cáncer de pulmón, por tanto, si quiere instalar un mecanismo de filtrado del aire, contacte la empresa ...".
- Criterio 3. El pasaje no cita fuentes expertas, pero no contiene desinformación ni anuncios o información de *marketing*. Es decir, la información que proporciona es correcta. Ejemplo: "El radón produce cáncer de pulmón".
- Criterio 4. El pasaje no cita fuentes expertas y el pasaje contiene anuncios o intenciones de *marketing*, pero sin contradecir el consenso médico general sobre el radón y sin proporcionar información incorrecta. Ejemplo: "El radón es una de las principales causas de cáncer de pulmón. Por eso es crucial disponer de medidores como el nuestro".
- Criterio 5. El pasaje no cita fuentes expertas, y contiene información incorrecta o que contradice el consejo médico, pero no contiene ni anuncios ni información de *marketing*. Ejemplo: "El radón no produce cáncer de pulmón".
- Criterio 6. El pasaje cita información proveniente de alguna de las siguientes procedencias: expertos/as reputados/as, artículos científicos, editoriales médicas, páginas de organismos gubernamentales, u otras fuentes similares que se estimen como autoritarias en la materia. Además, el pasaje contiene información incorrecta o que contradice el consejo médico, pero no contiene anuncios o intenciones de *marketing*. Ejemplo: "Según la OMS, el radón no es uno de los principales causantes de cáncer de pulmón", "Según el NHS, una exposición prolongada al gas radón puede producir diversos problemas de salud".
- Criterio 7. El pasaje no cita fuentes expertas y el pasaje contiene anuncios o intenciones de *marketing*, además contradice el consenso médico general sobre el radón y/o proporciona información incorrecta. Ejemplo: "El radón no es una de las principales causas de cáncer de pulmón. Por eso es crucial disponer de medidores como el nuestro".
- Criterio 8. El pasaje cita información proveniente de alguna de las siguientes procedencias: expertos/as reputados/as, artículos científicos, editoriales médicas, páginas de organismos gubernamentales, u otras fuentes similares que se estimen como autoritarias en la materia. Sin embargo, el pasaje contiene

anuncios o intenciones de *marketing*. Además, el pasaje contiene información incorrecta o que contradice el conocimiento médico. Ejemplo: "Recientes estudios de la EPA aseguran que la exposición a niveles altos de radón produce un impacto positivo en la salud de las personas para aliviar dolor articular entre otros, por tanto, visite nuestras instalaciones para llevar a cabo un tratamiento de radón."

El resultado de este etiquetado es un recurso de 2.056 pasajes extraídos de Internet y su estimación de calidad para consultas relacionadas con el radón. Los datos que obtuvimos de este etiquetado muestran que la mayoría de los pasajes, el **58,2 %**, pertenecen al nivel 2 de calidad. Por lo tanto, cumplirían el *Criterio 2* o *Criterio 3*, lo que quiere decir que es información correcta que o bien cita fuentes "autoritarias" y contiene intenciones de *marketing*, o bien no cita este tipo de fuentes, pero tampoco contiene intenciones de *marketing*. Por otra parte, el **33,4 %** de los pasajes se categorizan como de *Calidad 3*, es decir, la máxima calidad que cumple con el *Criterio 1*. Esto es, el pasaje cita información proveniente de alguna de las fuentes "autoritarias", además, el pasaje no contiene información que contraviene el consejo médico y tampoco contiene anuncios o intenciones de *marketing*. El **8,1 %** de los pasajes se enmarcarían en el nivel de *Calidad 1* y, por tanto, a pesar de no proporcionar información que contradice el consenso médico/científico, tendrían intenciones de marketing además de no citar fuentes expertas.

La cantidad de pasajes que proporcionarían información incorrecta (valores de calidad negativos), contradiciendo el consenso médico fue bajo, situándose en un **0,2 %**. Estos pasajes cumplen con alguno de los cuatro últimos criterios. En concreto, el **0,1 %** fue asociado a *Calidad -1*, el **0,04 %** a *Calidad -2* y el **0,09 %** a *Calidad -3*. Esto resalta que en el caso que nos atañe, los buscadores llevan a cabo un trabajo efectivo eliminando este tipo de páginas nocivas. Sin embargo, las páginas web detectadas que contenían información de la más baja calidad nos advierten de la amenaza que podría suponer para la salud pública si una gran cantidad de usuarios se topasen con ellas.

En algunos de estos casos, en el discurso del contenido de estas páginas se cuestiona y critica el consenso médico y científico. Ejemplo: *"Most people are not aware of the fact that there are actually no conclusive studies that have ever demonstrated that exposure to indoor radon, as commonly seen in the overwhelming vast majority of houses, increases the risk of cancer by any amount, and in fact, in the larger and better studies, what we see is that the risk of cancer actually goes down with increasing radon concentrations"*. En otras páginas webs se incita a los usuarios a llevar a cabo prácticas perjudiciales para su salud a través de un discurso basado en información falsa de forma intencionada con el fin de lucrarse económicamente. Ejemplo: "*The therapies can be both inhalation and immersion. The mechanisms for the beneficial effect of the radon spas covers a wide range of theories: from stimulation of the immune system, influences on homoeostasis, reduction in free radicals, and adaptive effects at the sub-cellular level on the genes, and DNA"*.

Por tanto, estos resultados ponen de manifiesto la necesidad de estar alerta para minimizar al máximo posible la presencia de estos contenidos en los *rankings* de resultados a través de la búsqueda web.

6. Discusión

En cuanto a los criterios de calidad de los pasajes, algunos anotadores percibieron la dimensión de "*intención de marketing*" con cierta ambigüedad. Una parte de los pasajes contenía información sobre datos de contacto de empresas o laboratorios que ofrecen servicios ajenos de medición o de mitigación del radón, pero que buscaban ofrecer medidas de apoyo a los usuarios. Sin embargo, a la vez, también están presentes pasajes en los que se ofrecen de forma explícita servicios propios de este tipo a través de anuncios comerciales con fines económicos. Ambos casos, se han considerado en la anotación como que contienen "*intenciones de marketing*" a pesar de que son casos distintos. Por ende, en pro de mejorar la precisión de estos criterios de etiquetación, y, por tanto, del recurso generado, sería conveniente en el futuro crear nuevas categorías dentro de las guías de etiquetación que contemplen este tipo de casos y los diferencien. Además, esto ha de hacerse de modo cuidadoso pues la complicación en exceso de los criterios de calidad puede afectar a la robustez del proceso de etiquetado.

Uno de los descubrimientos principales de este estudio es que unos criterios bien definidos conducen a etiquetas de mayor calidad y a un acuerdo mucho más sólido entre los evaluadores.

Aunque aún hay espacio para mejorar las directrices propuestas, hemos observado que, en nuestro experimento, incluso un breve proceso de instrucción a los asesores puede resultar en etiquetas más coherentes. Pese al número limitado de revisores y documentos anotados (algo inevitable en nuestro planteamiento al depender del factor humano como herramienta de trabajo), observamos señales prometedoras.

Una limitación de este estudio es que aún no podemos afirmar rotundamente que las directrices propuestas reflejan la calidad real de los documentos del corpus analizado. A pesar de que se ha demostrado que los criterios de relevancia de un pasaje son sólidos y robustos, el recurso final puede no ser tan preciso como sus criterios al haber quizás posibles sesgos en la anotación por parte de los etiquetadores. El acuerdo entre anotadores ha sido alto, pero los resultados finales podrían verse alterados si con las mismas directrices, usuarios sin experiencia previa o sin conocimientos de los riesgos del radón en la salud, la llevasen a cabo. Hay algunos aspectos específicos de la calidad intrínsecos al propio contenido que, al tratarse de información médica/científica, son difíciles de evaluar por etiquetadores no expertos en el campo. En el futuro, extenderemos esta investigación para abordar anotaciones con expertos del ámbito médico y compararemos los resultados y acuerdo con los obtenidos en el presente estudio.

7. Conclusiones

En esta investigación, hemos demostrado la dificultad de evaluar páginas web en términos de relevancia y calidad. Nuestra principal contribución es la creación de dos recursos de miles de páginas web etiquetadas en base a su relevancia y calidad para consultas relacionadas con los riesgos del radón para la salud. Estos recursos, como se ha mencionado anteriormente, pueden ser empleados en nuevos estudios en una gran variedad de campos con particular énfasis en el desarrollo de soluciones orientadas a la detección de desinformación. Además, otro de los resultados a destacar es la creación de un conjunto de pautas para crear anotaciones sólidas que pueden mejorarse aún más mediante una breve capacitación a los evaluadores. En trabajos futuros, tenemos la intención de seguir puliendo estas pautas y llevar a cabo un estudio con usuarios para comprender en más detalle cómo la información de baja y alta calidad es percibida por usuarios reales de sistemas de búsqueda. Esperamos que tanto los recursos de etiquetado generados como los criterios y pautas de anotación creadas tengan un impacto positivo y sean empleados en nuevas investigaciones.

8. Agradecimientos

Esta publicación es parte/cuenta con la financiación de los siguientes proyectos de I+D+i: Este artículo se elaboró en el marco del proyecto Radón en España: percepción de la opinión pública, agenda mediática y comunicación del riesgo (RAPAC) del Consejo de Seguridad Nuclear (SUBV-13/2021) & el proyecto Medios nativos digitales en España: estrategias, competencias, implicación social y (re)definición de prácticas de producción y difusión periodísticas (PID2021-122534OB-C21), financiado por MCIN/ AEI/10.13039/501100011033/ y "FEDER Una manera de hacer Europa". Este trabajo ha sido financiado por el proyecto PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU). Los autores agradecen también el apoyo financiero prestado por la Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidade (ED431G 2023/04, ED431C 2022/19) y al Fondo Europeo de Desarrollo Regional, que reconoce al CiTIUS-Centro de Investigación en Tecnologías Inteligentes de la Universidad de Santiago de Compostela como Centro de Investigación del Sistema Universitario de Galicia. David E. Losada agradece el apoyo financiero obtenido del proyecto SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego) y del proyecto PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; apoyado por el Fondo Europeo de Desarrollo Regional).

9. Referencias

[1] K. Schwab, "La cuarta revolución industrial," *Futuro Hoy*, vol. 1, no. 1, pp. 6-10, 2020. [Online]. Available: https://bit.ly/Schwabrevistafuturo

[2] L. J. F. Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser, y B. W. Hesse, "Online health information seeking among US adults," *Public Health Reports*, vol. 134, no. 6, pp. 617-625, 2019, doi: 10.1177/0033354919874074.

[3] S. S. Tan y N. Goonawardene, "Internet health information seeking and the patient-physician relationship: A systematic review," *Journal of Medical Internet Research*, vol. 19, no. 1, e9, 2017, doi: 10.2196/jmir.5729.

[4] B. Osei Asibey, S. Agyemang, y A. Boakye Dankwah, "The internet use for health information seeking among Ghanaian university students: A cross-sectional study," *International Journal of Telemedicine and Applications*, 1756473-9, 2017, doi: 10.1155/2017/1756473.

[5] B. Swire-Thompson y D. Lazer, "Public health and online misinformation: Challenges and recommendations," *Annual Review of Public Health*, vol. 41, no. 1, pp. 433-451, 2020, doi: 10.1146/annurev-publhealth-040119-094127.

[6] G. Eysenbach, "Infodemiology: The epidemiology of (mis)information," *The American Journal of Medicine*, vol. 113, no. 9, pp. 763-765, 2002, doi: 10.1016/s0002-9343(02)01473-0.

[7] F. A. Pogacar, A. Ghenai, M. D. Smucker, y C. L. A. Clarke, "The positive and negative influence of search results on people's decisions about the efficacy of medical treatments," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 209-216, 2017, doi: 10.1145/3121050.3121074.

[8] S. Jiang y P. L. Liu, "Digital divide and internet health information seeking among cancer survivors: A trend analysis from 2011 to 2017," *Psycho-Oncology*, vol. 29, no. 1, pp. 61-67, 2020, doi: 10.1002/pon.5247.

[9] A. Montoro-Montarroso, J. Cantón-Correa, P. Rosso, B. Chulvi, Á. Panizo-Lledot, J. Huertas-Tato, B. Calvo-Figueras, M. J. Rementeria y J. Gómez-Romero, "Fighting disinformation with artificial intelligence: fundamentals, advances and challenges," *Profesional de la Información*, vol. 32, no. 3, e320322, 2023, doi: 10.3145/epi.2023.may.22.

[10] A M. Fernández-Pichel, S. Meyer, M. Bink, A. Frummet, D. E. Losada, and D. Elsweiler, "Improving the reliability of health information credibility assessments," in *Proc. ROMCIR*, 2023.

[11] D. Zhu, S. L. Nimmagadda, K. W. Wong, y T. Reiners, "Relevance Judgment Convergence Degree—A Measure of Assessors Inconsistency for Information Retrieval Datasets," en *International Conference on Information Systems Development*, pp. 149-168, Cham, Switzerland: Springer International Publishing, 2022.

[12] F. L. Cruz, J. A. Troyano, F. Enríquez, and F. J. Ortega, "Detección y clasificación de falacias prototípicas y espontáneas en español," *Procesamiento del Lenguaje Natural*, vol. 71, pp. 53-62, 2023.

[13] A. Nabożny, B. Balcerzak, A. Wierzbicki, M. Morzy, y M. Chlabicz, "Active annotation in evaluating the credibility of Web-based medical information: Guidelines for creating training data sets for machine learning," *JMIR Medical Informatics*, vol. 9, no. 11, e26065, 2021

[14] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, y A. X. Mina, "A structured response to misinformation: Defining and annotating credibility indicators in news articles," in *Companion Proceedings of The Web Conference 2018*, pp. 603-612, April 2018.

[15] OMS, "El radón y sus efectos en la salud," [Online]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/radon-and-health, 2021.

[16] J. M. Samet, "Radon and lung cancer," *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 10, pp. 745-758, 1989, doi: 10.1093/jnci/81.10.745.

[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, P. J. Liu y otros, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485-5551, 2020.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, vol. 463, New York, NY, USA: ACM Press, 1999.

[19] W. B. Croft, D. Metzler, and T. Strohman, “Search Engines: Information Retrieval in Practice”, vol. 520, Reading: Addison-Wesley, 2010, pp. 131-141.

[20] S. Robertson y H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond,"

Foundations and Trends® in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009.

[21] R. Nogueira, Z. Jiang, y J. Lin, "Document ranking with a pretrained sequence-to-sequence model," arXiv preprint arXiv:2003.06713, 2020.

[22] C. L. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, and G. Zuccon, "Overview of the TREC 2020 Health Misinformation Track," in *TREC*, Nov. 2020.