MONOGRÁFICO

Documentación de las **Ciencias** de la **Información**
ISSN-e: 1988-2890

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/dcin.77472

Modelos algorítmicos y *fact-checking* automatizado. Revisión sistemática de la literatura[1]

David García-Marín[2]

Recibido: 26 de julio de 2021 / Aceptado: 8 de noviembre de 2021

Resumen. El *fact-checking* automatizado consiste en la comprobación automática de la veracidad de una información aplicando las tecnologías de inteligencia artificial existentes para clasificarla en alguna de las categorías comúnmente usadas por los *fact-checkers* humanos (verdadero, engañoso, falso, etc.). Este trabajo presenta el primer análisis bibliométrico en castellano -de tipo cuantitativo- sobre la evolución y los países de procedencia de la investigación sobre esta práctica. Asimismo, pretende analizar el nivel de precisión de las soluciones algorítmicas y el impacto de los trabajos publicados, utilizando para ello tratamientos estadísticos descriptivos e inferenciales (pruebas de chi cuadrado y test de Kruskal-Wallis). De acuerdo con nuestros resultados, en los últimos tres años se concentra el mayor volumen de aportaciones, que proceden mayoritariamente de la región asiática y Estados Unidos. Predominan los trabajos que proponen métodos o sistemas algorítmicos integrados. Son mayoritarios los estudios sobre modelos lingüísticos, que presentan aún varias limitaciones y una efectividad inferior a la media. Se observa una reducida atención hacia los modelos basados en el análisis de imágenes, y resulta prácticamente nula la presencia de algoritmos de detección de audios falsos. En línea con trabajos anteriores, nuestro estudio concluye que no existen diferencias estadísticamente significativas en el nivel de precisión de los diversos modelos algorítmicos propuestos, a pesar de sus diferentes grados de complejidad técnica.
Palabras clave: *fact-checking* automatizado; desinformación; *fake news*; revisión sistemática; algoritmos

[en] Algorithmic models and automated fact-checking. A systematic literature review

Abstract. Automated fact-checking consists of automatically determining the veracity of a claim by applying existing artificial intelligence technologies to classify it into one of the categories commonly used by human fact-checkers (true, misleading, false, etc.). This paper presents the first systematic literature review in Spanish on the evolution of research on this topic. It also aims to analyze the level of accuracy of algorithmic solutions and the impact of published work, using descriptive and inferential statistical treatments (chi-square and Kruskal-Wallis tests). According to our results, the highest volume of contributions was concentrated in the last three years, mainly from the Asian region and United States. Papers proposing integrated algorithmic methods or systems predominate. Studies on linguistic models, which still have several limitations and below-average effectiveness, are in the majority. There is little attention to models based on image analysis, and the presence of fake audio detection algorithms is practically nonexistent. In line with previous work, our study concludes that there are no statistically significant differences in the level of accuracy of the diverse algorithmic models proposed, despite their different degrees of technical complexity.
Keywords: automated fact-checking; disinformation; fake news; systematic review; algorithms

Cómo citar: García-Marín, D. (2022): Modelos algorítmicos y *fact-checking* automatizado. Revisión sistemática de la literatura, en *Documentación de Ciencias de la Información* 45 (1), 7-16.

1. Introducción y marco teórico

El fenómeno de la desinformación es uno de los principales objetos de estudio en el campo de la comunicación desde su eclosión en su forma actual a partir de 2016. Aunque este fenómeno no resulta novedoso y ha estado presente en todas las etapas históricas, el proceso electoral estadounidense que terminó con la llegada de Donald Trump a la Casa Blanca y el Brexit en el Reino Unido marcaron el inicio de la atención mediática y académica hacia este desafío. Desde entonces, se han popularizado conceptos como desinformación, posverdad y *fake news* que, aunque están vinculados entre sí, presentan diferentes matices conceptuales. Más que un tipo de contenido concreto, la desinformación es un proceso donde las noticias falsas se mezclan con información verídica y donde diferentes canales cooperan para potenciar la credibi-

[1] Trabajo apoyado por la Cátedra Jean Monnet "EUDFAKE: EU, disinformation and fake news" financiada por el programa Erasmus + de la Comisión Europea. También cuenta con el apoyo del proyecto "Racionalidad y contraconocimiento. Epistemología de la detección de falsedades en relatos informativos" financiado por el Ministerio de Ciencia, Innovación y Universidades de España.

[2] Universidad Rey Juan Carlos (España)
E-mail: david.garciam@urjc.es
ORCID: https://orcid.org/0000-0002-4575-1911

lidad de tales relatos maliciosos (Elías, 2021). En el mismo sentido, la posverdad se relaciona con aquellos procesos donde la realidad se explica desde lo emocional en lugar de atender a los hechos empíricamente demostrados. En este contexto, las *fake news* son una manifestación concreta de la desinformación consistente en falsedades elaboradas con una apariencia que emula el estándar periodístico (McIntyre, 2018), por lo que un amplio volumen de ciudadanos puede decodificar tales mensajes como verídicos.

Una de las principales tendencias de la investigación sobre esta temática se enmarca en la búsqueda de soluciones para afrontar este desafío. Los trabajos que indagan en estos posibles abordajes se centran en el refuerzo de los valores clásicos del periodismo para recuperar la credibilidad perdida como garantes de una información veraz (Aparici y García-Marín, 2019; Ruiz-Rico, 2020) que realice una efectiva curación de contenidos (Aleixandre-Benavent, Castelló-Cogollos y Valderrama-Zurián, 2020) y ofrezca la información basada en el interés general que los públicos demandan. Asimismo, han sido numerosos los trabajos centrados en la necesaria alfabetización informacional de la ciudadanía a fin de diferenciar la informacion maliciosa de los contenidos verídicos (Caldevilla-Domínguez y García-García, 2020; García-Ortega y García-Avilés, 2021; Núñez-Mussa, 2020) a partir de la implantación de proyectos formativos formales y no formales tanto dentro como fuera de las instituciones educativas.

En la misma línea, un volumen relevante de textos analiza el periodismo de verificación o *fact*-checking como herramienta central en la lucha contra las *fake news*. Estas entidades realizan una labor de comprobación de la veracidad de los contenidos a posteriori, después de su emisión o publicación. Estos *fact-checkers* fundamentan su tarea de verificación en los siguientes aspectos:

- El conocimiento empírico. Bajo esta aproximación, se pretende determinar si el contenido a verificar está soportado por hechos demostrados (Magdy y Wanas, 2010; Lease, 2018).
- El contexto. Se establece la veracidad de una historia analizando sus metadatos, como la credibilidad de su autor o la velocidad y forma de diseminación en las redes sociales (Antonakaki, Fragopoulou y Ioannidis, 2021).
- El contenido. Se determina la veracidad de una historia a partir de sus marcas textuales, visuales o sonoras basándose en consideraciones de estilo como, por ejemplo, la extensión y variedad de las palabras o la complejidad sintáctica del texto (Horne y Adali, 2017; Potthast et al., 2018; Oshikawa, Qian y Wang, 2020).

Dado el volumen creciente de producción de contenidos desinformativos, en los últimos años ha comenzado la experimentación con sistemas de inteligencia artificial para detectar noticias falsas, a fin de facilitar la labor de estas entidades. El *fact-checking* automatizado o computacional consiste, por tanto, en la comprobación automática de la veracidad de una información aplicando las tecnologías de inteligencia artificial existentes para clasificarla en alguna de las categorías comúnmente usadas por los *fact-checkers* humanos: verdad, media verdad, engañosa, falsa, etc. (Dale, 2017; Saquete et al., 2019). Los algoritmos o sistemas algorítmicos empleados en la detección automática de noticias falsas pueden adoptar los siguientes modelos, propuestos por Choras et al. (2020):

1. Lingüísticos. Los más básicos consisten en la contabilización de la frecuencia de aparición de determinadas palabras o secuencias de palabras dentro del texto a analizar. En concreto, estudian la existencia de determinadas claves lingüísticas y psicolingüísticas para detectar contenidos falsos, tales como (1) la falta de diversidad en el uso de construcciones gramaticales, (2) la presencia de redundancias, (3) la ausencia de complejidad en el lenguaje (frases incompletas, textos no estructurados, vocabulario limitado, etc.), (4) el uso de expresiones informales e inciertas, (5) los errores gramaticales, ortográficos o de concordancia, (6) la falta de información de contexto, (7) la utilización de expresiones emocionales y (8) el uso frecuente de adjetivos y adverbios (Saquete et al., 2019). Como observamos, estos métodos no solo se basan en el análisis de las secuencias de palabras y las emociones del autor escondidas en el texto, sino que deben atender también a la estructura sintáctica de las frases. Para entrenar a los algoritmos, resulta fundamental la elaboración de bases de datos específicas para cada idioma. Asimismo, en estos modelos basados en el análisis textual, son muy útiles las bases de datos constituidas por grandes cantidades de afirmaciones etiquetadas como verdaderas o falsas (y sus diferentes matices: engañosa, media verdad, etc.), procedentes del trabajo de las entidades de verificación. Todas estas bases necesitan organizar sus datos de forma estructurada para que puedan funcionar correctamente en modelos de aprendizaje automático (*machine learning*), como explicaremos más adelante.
2. Reputacionales. Su objetivo es evaluar el grado de credibilidad de la fuente que emite una determinada información. Para establecer este nivel de confianza, estos algoritmos utilizan, fundamentalmente, dos tipos de datos: (1) la reputación del contenido general que publica esta fuente, junto con las revisiones / *feedback* de terceros a propósito de este contenido, y (2) la medición de su reputación a partir del análisis de su IP o su dominio online.
3. Análisis de redes. Si los dos anteriores modelos se basan en claves lingüísticas de los textos y en el análisis de la reputación de sus creadores, los

algoritmos de análisis de redes evalúan la credibilidad de un contenido analizando (1) cómo se propaga, (2) quiénes son sus difusores y (3) qué relaciones existen entre ellos dentro de una red.

4. Análisis basado en imágenes. Estos modelos son capaces de reconocer si una imagen ha sido alterada. Para ello, analizan las huellas que cada modelo de cámara deja en las imágenes, a fin de detectar si una fotografía o vídeo han sido compuestos por fragmentos captados con diferentes dispositivos.

Para realizar la detección automática de los contenidos falsos o fuentes maliciosas, estos modelos algorítmicos pueden utilizar procedimientos de aprendizaje automático (*machine learning*) o aprendizaje profundo (*deep learning*). El primero se basa en un aprendizaje supervisado donde los algoritmos son entrenados utilizando ejemplos etiquetados que expresan cuáles son las respuestas correctas. El algoritmo recibe un conjunto de instrucciones, realiza sus predicciones y compara sus resultados con las respuestas válidas, previamente provistas por el diseñador. Al comparar sus respuestas con las correctas, encuentra sus errores y modifica su funcionamiento convenientemente.

El aprendizaje profundo es una modalidad más compleja y avanzada de aprendizaje automático que emula metafóricamente el funcionamiento de las redes neuronales del cerebro humano. Se fundamenta en el funcionamiento en red de varios algoritmos de aprendizaje automático que actúan en varias capas para producir un conocimiento más complejo sin necesidad de un conjunto de datos estructurado ni de ejemplos etiquetados (no reciben información sobre las respuestas correctas), ya que son capaces por sí mismos de elaborar patrones y buscar soluciones (Ongsulee, 2017).

En este contexto inicial de la investigación sobre las soluciones tecnológicas contra el fenómeno de la desinformación, este trabajo ofrece una primera aproximación en nuestro idioma a la producción de algoritmos para el periodismo de verificación automatizado. El objetivo es presentar un análisis bibliométrico de tipo cuantitativo sobre la evolución y la procedencia de la investigación en esta práctica. Se pretende analizar también su nivel de precisión, así como el impacto de las investigaciones que han abordado este objeto de estudio. La finalidad última es producir un estado de la cuestión de las tecnologías algorítmicas aplicadas a la verificación periodística de hechos.

2. Preguntas de investigación y método

Para alcanzar el objetivo anteriormente mencionado, se llevó a cabo una revisión sistemática de la producción científica sobre *fact-checking* automatizado, centrada en las propuestas algorítmicas empíricamente validadas para detectar contenidos falsos. En concreto, se plantean las siguientes preguntas de investigación:

1. PI1. ¿Cuáles son los años de publicación, el país de procedencia, los tipos de propuesta y los modelos algorítmicos de los trabajos sobre *fact-checking* automatizado publicados?
2. PI2. ¿Cuál es el nivel de precisión de cada modelo algorítmico, de acuerdo con la evidencia empírica aportada por estos trabajos? ¿Existen diferencias significativas en el nivel de precisión de cada uno de estos modelos?
3. PI3. ¿Cuál es el impacto de estos estudios (medido en número de citas recibidas) en función del modelo algorítmico propuesto? ¿Existen diferencias significativas en el impacto de estos trabajos en función de su modelo algorítmico?

Obsérvese que las tres preguntas de investigación integran los tres niveles de lectura de análisis bibliométrico, de acuerdo con Polanco (1997) y Puebla-Martínez, Del Campo y Pérez-Cuadrado (2018): (1) el bibliográfico o relativo a las revistas (año de publicación), (2) el sociológico o relativo a los autores (país de procedencia) y (3) el conocimiento objetivo o relativo a los textos (tipo de propuesta, modelo algorítmico, eficacia e impacto).

Para conformar la muestra de trabajos a analizar, fueron incluidos los artículos de investigación evaluados por pares que presentan propuestas concretas de *fact-checking* automatizado, independientemente de su modelo algorítmico y de su grado de fiabilidad demostrado. Se incluyeron tanto artículos publicados en su versión definitiva como documentos en acceso anticipado (*early access*). Se decidió excluir los siguientes tipos de trabajo: (1) conferencias y actas de congresos (*proceedings*), (2) textos no alineados con el enfoque temático anteriormente descrito y (3) aquellos que, centrados en el *fact-checking* automatizado, no presentan una propuesta concreta para la detección de *fake news*, tales como revisiones de la literatura, ensayos y artículos de corte teórico, estudios sobre las percepciones hacia esta práctica, editoriales de presentación de números monográficos y, finalmente, estudios comparativos de la eficacia de sistemas computacionales y algoritmos ya existentes.

La búsqueda se realizó en la base de datos Web of Science (repositorio Core Collection en todas sus ediciones). Trabajos anteriores similares al nuestro, como el de López García et al. (2019), han utilizado exclusivamente esta base de datos por ser la de mayor relevancia a nivel internacional y recoger las revistas de mayor visibilidad e impacto.

La búsqueda documental se realizó a partir de varias cadenas de conceptos clave tomando como referencia el título, el resumen y las *keywords* de los artículos publicados hasta el 31 de mayo de 2021, sin establecimiento de fecha inicial. La tabla 1 recoge los operadores *booleanos* utilizados y el número de artículos recuperados en las diferentes fases de filtrado. Tras una primera búsqueda, se recogieron un total de 670 artículos, que se sometieron a un flujo de trabajo dividido en 3 fases:

1. Revisión del tipo de documento. Se eliminaron los tipos de trabajo sujetos a exclusión anteriormente expuestos (n=373). El *corpus* quedó reducido a 297 textos.
2. Lectura del título, resumen y palabras clave (y revisión del texto completo en caso de duda). Fueron excluidos 113 trabajos por situarse fuera del foco de la investigación, considerándose un total de 164 artículos.
3. Eliminación de los textos duplicados (n=68) relativos a los distintos términos de búsqueda empleados. La muestra final quedó establecida en 96 trabajos (listado completo de artículos, disponible en: https://cutt.ly/GQq3Zjl).

Tabla 1. Términos de búsqueda y número de trabajos recuperados por fase.

Términos de búsqueda	Búsqueda incial	Fase 1	Fase 2	Fase 3
Automat* AND fact-checking	92	33	18	7
Automat* AND detection AND fake news	163	72	54	27
Automat* AND detection AND disinformation	26	13	8	5
Automat* AND detection AND misinformation	99	44	15	12
Automat* AND detection AND post-truth	4	4	0	0
Algorithm* AND fact-checking	74	40	19	5
Algorithm* AND detection AND fake news	117	53	27	25
Algorithm* AND detection AND disinformation	17	10	8	4
Algorithm* AND detection AND misinformation	75	27	15	11
Algorithm* AND detection AND post-truth	3	1	0	0
Total	670	297	164	96

Tras la construcción de la muestra, se procedió a elaborar una ficha de registro de la información, conformada por las variables presentes en las preguntas de investigación:

1. Año de publicación de los trabajos.
2. País de procedencia. En el caso de artículos firmados por investigadores de varios países, se tomó como referencia el país del autor de correspondencia.
3. Tipo de propuesta. Las posibles aportaciones que integran los artículos seleccionados son: (1) bases de datos en cualquier idioma (o en varios) para la posterior construcción de modelos de detección lingüísticos o audiovisusales, (2) algoritmos específicos encaminados a la clasificación del contenido desinformativo, y (3) sistemas / métodos algorítmicos completos de detección de desinformación que combinan más de un algoritmo e, incluso, incorporan su propia base de datos.
4. Modelo algorítmico. Se utilizó una adaptación de la clasificación de Choras et al. (2020), explicada en el apartado anterior. Se contemplaron, por tanto, los siguientes posibles modelos: (1) lingüístico, (2) reputacional, (3) análisis de redes, (4) análisis de imágenes, (5) análisis de audio, y (6) análisis mixto (aquellos que combinan, al menos, dos de los modelos anteriores).
5. Nivel de precisión de la propuesta. Se asume el dato de precisión global del algoritmo o del sistema presentado de acuerdo con las pruebas empíricas realizadas por los propios autores, recogidas en sus trabajos. Para esta variable, se toma el valor F, que consiste en la media armónica de la tasa de efectividad y la tasa de precisión (Thaer et al. 2021). La tasa de efectividad mide el número de contenidos falsos correctamente detectados sobre el total de contenidos falsos que forman parte de la comprobación. La tasa de precisión es el número de contenidos falsos correctamente detectados sobre el número total de predicciones realizadas. En caso de probarse varios modelos, se tomó el valor de precisión más elevado. Del mismo modo, se asumió el valor más alto en caso de probarse el mismo modelo varias veces.
6. Número de citas recibidas (hasta el 31 de mayo de 2021).

Los datos recogidos en la ficha de registro fueron tratados con procedimientos estadísticos descriptivos e inferenciales utilizando el software SPSS v.26. Se recurrió a la estadística inferencial como comple-

mento a la información descriptiva con el fin de confirmar la existencia de diferencias significativas entre las diferentes categorías de las variables. Aunque en ocasiones estas diferencias pueden parecer obvias a simple vista, las pruebas de tipo inferencial ofrecen una visión más precisa sobre su nivel de significatividad. Este nivel, que puede tener diferentes rangos (desde no existir hasta tener valores muy elevados), es medido numéricamente por estas pruebas. Asimismo, en determinadas circunstancias resulta muy complicado determinar si tales diferencias resultan significativas, como sucede en este trabajo en el caso de la efectividad de los modelos algorítmicos analizados y el impacto de los artículos que los describen, como veremos en el siguiente apartado. Por ello, es necesario acudir a pruebas de este tipo para conocer si el tipo de modelo algorítmico constituye un factor estadísticamente significativo en (1) la efectividad de detección de contenidos falsos y (2) el impacto de los artículos donde tales modelos son descritos. Por último, estas pruebas nos permiten comparar niveles de significatividad de diferentes variables. ¿Son mayores las diferencias entre trabajos realizados por países o por año de publicación? Para obtener esta información, necesitamos cuantificar el grado de significatividad ya que con la mera observación de los datos resulta imposible ofrecer una respuesta.

En primer lugar, se ejecutaron pruebas de normalidad mediante test de Kolmogorov-Smirnov con corrección de significación de Lilliefors a fin de decidir la aplicación de procedimientos paramétricos o no paramétricos para comprobar la existencia de diferencias significativas entre las categorías de las variables. Este test observó la ausencia de distribuciones normales en todas las variables (tabla 2), por lo que se realizaron pruebas no paramétricas mediante test de chi cuadrado. Además, se comprobó si el modelo algorítmico presentado en cada trabajo constituye una variable determinante de (1) el nivel de precisión de la propuesta (PI2) y (2) el impacto del trabajo medido en número de citas (PI3). Para ello, se recurrió a la prueba no paramétrica de Kruskal-Wallis.

Tabla 2. Pruebas de normalidad (Kolmogorov-Smirnov). *Se establecen diferencias muy significativas cuando $p < 0,01$

Variable	Estadístico de prueba	p valor
Año	0,233	0,000*
País	0,118	0,002*
Propuesta	0,401	0,000*
Modelo algorítmico	0,343	0,000*

3. Resultados

3.1. Años de publicación, países, propuestas y modelos algorítmicos

Los trabajos que forman parte de la muestra se concentran en los tres últimos años (tabla 3). Entre 2019 y 2021 se sitúa el 83,33% de los textos, que experimentan un creciente ritmo de publicación. A pesar de no ser computado completo ya que nuestro trabajo solo recogió manuscritos publicados hasta el 31 de mayo, el año 2021 presenta el mayor número de publicaciones (n=37; 38,54%). Las pruebas de chi cuadrado confirman la existencia de diferencias muy significativas en el número de trabajos realizados cada año: $\chi 2(10, N = 96) = 177.62, p = 0.000$.

Tabla 3. Años de publicación de los artículos analizados.

Año de publicación	Frecuencias
2008	1 (1,04%)
2009	1 (1,04%)
2013	1 (1,04%)
2014	1 (1,04%)
2015	2 (2,08%)
2016	1 (1,04%)
2017	3 (3,12%)
2018	6 (6,25%)
2019	17 (17,70%)
2020	26 (27,08%)
2021	37 (38,54%)
Total	96 (100%)

Estados Unidos (n=18; 18,75%), India (n=14; 14,58%) y China (n=12; 12,50%) son los tres países con mayor número de estudios publicados, que representan cerca de la mitad de los textos de la muestra (45,83%) (tabla 4). Es destacable también la producción de trabajos desde Grecia (n=6; 6,25%) e Italia (n=5; 5,20%). Un total de 27 países han publicado, al menos, un artículo. Resulta llamativo el elevado número de países asiáticos (n=12) dedicados a la creación y publicación de algoritmos para la detección de contenidos falsos. Un total de 39 textos (40,62%) proceden de Asia, cifra que supera el volumen de propuestas realizadas desde el continente americano (n=29; 30,20%) y Europa (n=26; 27,08%) (distribución por países completa, disponible en el siguiente enlace: https://cutt.ly/jQq4xeY). Las diferencias en el número de trabajos publicados en función del país resultan muy significativas [$\chi 2(26, N = 96) = 138.00, p = 0.000$].

Tabla 4. Países de procedencia. Se recogen los países que registran más de 3 trabajos.

País	Frecuencias
Estados Unidos	18 (18,75%)
India	14 (14,58%)
China	12 (12,50%)
Grecia	6 (6,25%)
Italia	5 (5,20%)
España	4 (4,16%)
Brasil	4 (4,16%)
Canadá	4 (4,16%)
Turquía	4 (4,16%)

También se aprecian diferencias muy relevantes en el tipo de propuestas presentadas [$\chi 2(2, N = 96) = 140.75, p = 0.000$]. Existe un predominio de los trabajos que proponen sistemas o métodos completos para la detección de contenidos desinformativos (n=74; 77,08%), frente a la menor cantidad de estudios sobre bases de datos (n=11; 11,45%) y algoritmos específicos (n=9; 9,37%) (tabla 5). En el caso de las bases de datos, son mayoritarias las realizadas en lengua inglesa con el fin de entrenar algoritmos utilizados en modelos de análisis lingüístico en este idioma; si bien existen aproximaciones centradas en otros, como el portugués (Silva, Santos, Almeida y Pardo, 2020), el español (Posadas-Durán, Gómez-Adorno, Sidorov y Moreno-Escobar, 2019) y lenguas minoritarias como el urdu (Amjad et al. 2020).

Los modelos algorítmicos basados en el análisis lingüístico resultan predominantes (n=53; 55,20%) (tabla 5). Tras ellos, se sitúan los mixtos (n=23; 23,95%), que atienden, básicamente, a las siguientes combinaciones: (1) análisis lingüístico + análisis de redes y (2) análisis lingüístico + análisis reputacional. Por el contrario, las propuestas algorítmicas centradas exclusivamente en el análisis de imágenes resultan minoritarias (n=4; 4,16%), y las basadas en el audio son prácticamente inexistentes (n=1; 1,04%). Las diferencias en las frecuencias de esta variable resultan también muy significativas [$\chi 2(5, N = 96) = 122.25, p = 0.000$].

Tabla 5. Tipos de propuesta y modelos algorítmicos.

Propuestas	
Algoritmo	9 (9,37%)
Método / sistema	76 (79,16%)
Base de datos	11 (11,45%)
Modelos algorítmicos	
Lingüístico	53 (55,20%)
Reputacional	4 (4,16%)
De redes	11 (11,45%)
De imágenes	4 (4,16%)
Mixto	23 (23,95%)
De audio	1 (1,04%)

3.2. Nivel de precisión

Un total de 62 trabajos que proponen algoritmos o sistemas algorítmicos para detectar *fake news* presentan estudios empíricos sobre su precisión. Esta cifra constituye el 72,94% de los 85 estudios que realizan este tipo de propuestas (se excluyen los 11 trabajos que presentan bases de datos). Se han detectado 23 textos que no ofrecen valores generales sobre el nivel de precisión de su sistema / algoritmo, bien porque tales pruebas no fueron realizadas (al menos, no figuran en el manuscrito) o porque, tras ejecutarse, el dato que el artículo aporta es un valor relativo frente a los de otras propuestas, no una tasa de eficacia absoluta.

El nivel medio de precisión de estas 62 soluciones algorítmicas se sitúa en el 89,81% (valor F). Los análisis de redes registran el índice más elevado (M=94,53%), seguidos de los análisis de imágenes (M=92,87%) y los mixtos (M=92,78%) (tabla 6). El único trabajo centrado en el audio presenta un índice de efectividad marcadamente inferior al resto (M=74,10). Este hecho puede deberse a la reducida atención prestada a la desinformación en este formato, frente a la mayor cantidad de estudios centrados en el texto y la imagen. Las pruebas no paramétricas mediante test de Kruskal-Wallis no observan diferencias significativas en el nivel de precisión en función del tipo de modelo algorítmico utilizado [$H(5) = 6.927, p = 0.226$] (figura 1).

Tabla 6. Nivel de precisión de cada modelo algorítmico.

Modelo algorítmico	n	Promedio de precisión (%)	*D.T.*
Lingúístico	34	88,23	11,01
Reputacional	3	87,36	6,18
De redes	3	94,53	1,33
Mixto	18	92,78	7,30
De imágenes	3	92,87	8,04
De audio	1	74,10	-
Total	**62**	**89,81**	**9,68**

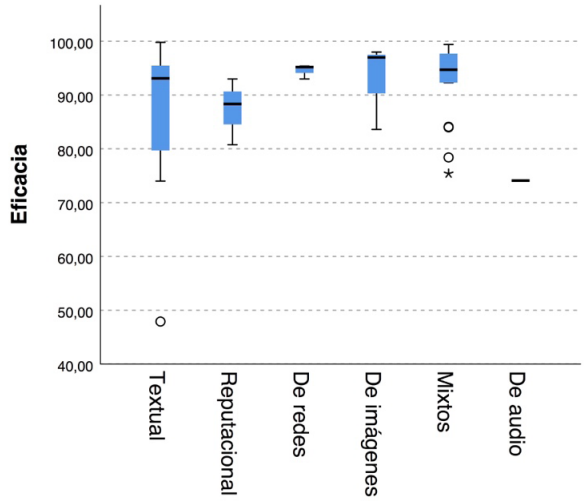

Figura 1. Gráfico de cajas del nivel de precisión.

La tabla 7 recoge los 10 trabajos con mayor nivel de precisión, de acuerdo con las pruebas realizadas por los propios autores.

Tabla 7. Propuestas con mayor nivel de precisión.

Autor/es	Modelo algorítmico	Precisión (%)
Goldani, Momtazi y Safabakhsh (2021)	Lingüístico *Deep learning*	99,80
Huang y Chen (2020)	Mixto (lingüístico + redes) *Deep learning*	99,40
Sahoo y Gupta (2021)	Mixto (lingüístico + reputacional) *Deep learning*	99,40
Zervopoulos et al. (2021)	Mixto (lingüístico + redes) *Deep learning*	99,30
Ilias y Roussaki (2021)	Lingüístico *Deep learning*	99,01
Mouratidis, Nikiforos y Kermanidis (2021)	Mixto (lingüístico + redes) *Deep learning*	99,00
Kaur, Kumar y Kumaraguru (2019)	Lingüístico *Machine learning*	98,80
Gereme et al. (2021)	Lingüístico *Deep learning*	98,53
Goldani, Safabakhsh y Momtazi (2021)	Lingüístico *Deep learning*	98,12
Kasban y Nassar (2020)	De imágenes *Deep learning*	98,00

3.3. Impacto de los trabajos

Por último, resulta llamativo el reducido número de citas que reciben los estudios que conforman la muestra (M=7,21), a pesar de ser publicados en revistas de primer nivel. Esta baja repercusión puede deberse a su reciente aparición y a la escasez de investigadores que todavía presenta este novedoso campo, carente aún de una comunidad vigorosa. Los modelos de análisis de imágenes consiguen, con gran diferencia, un impacto mayor (M=23,67), seguidos de los lingüísticos (M=8,94) y los de redes (M=6,27) (tabla 8). A pesar de estas diferencias, el modelo algorítmico presentado tampoco constituye una variable determinante en el impacto de estos trabajos [$H(5) = 5.225$, $p = 0.389$] (figura 2).

Tabla 8. Impacto de los artículos en función del modelo algorítmico.

Modelo algorítmico	n	Promedio de citas recibidas	*D.T.*
Lingüístico	54	8,94	18,13
Reputacional	4	4,00	7,34
De redes	11	6,27	7,73
De imágenes	3	23,67	40,13
Mixto	23	2,48	4,44
De audio	1	0	-
Total	**96**	**7,21**	**15,64**

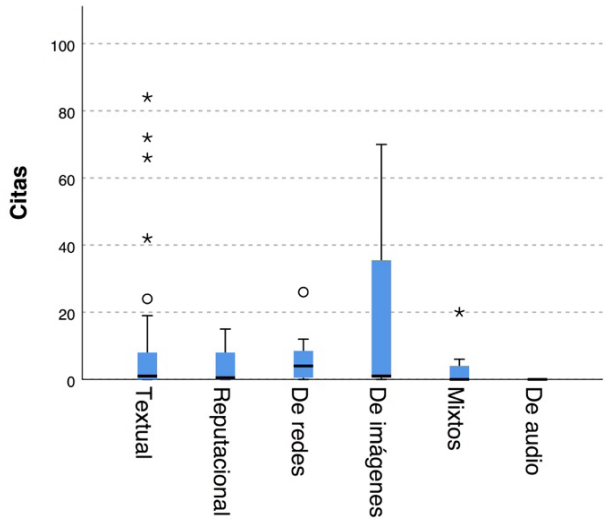

Figura 2. Gráfico de cajas del número de citas recibidas.

4. Discusión y conclusiones

La aplicación de soluciones algorítmicas para automatizar la detección de información falsa ha experimentado una tendencia expansiva en el ámbito investigador a partir de 2019. En los últimos tres años se concentra el mayor volumen de aportaciones, que presentan un ritmo creciente de publicación. La gran mayoría de los textos de nuestra muestra (8 de cada 10) han sido publicados en los tres últimos años, y el 38,54% ha salido a la luz en la primera mitad de 2021 (PI1). La novedad de este objeto de estudio determina el aún bajo impacto de este tipo de trabajos (medido en el número de citas recibidas) (PI3), lo que nos sitúa ante un campo de conocimiento todavía en un estado ciertamente liminar, aunque en clara expansión y "con gran potencial de explotación" (Meneses-Silva, Silva-Fontes y Colaço-Júnior, 2021, p. 185).

Estados Unidos, India y China son los países más prolíficos en este tipo de propuestas (PI1), confirmando así los resultados de estudios anteriores como el de Meneses-Silva, Silva-Fontes y Colaço-Júnior (2021). Resulta destacable la elevada presencia de países asiáticos, región que lidera la investigación aplicada de soluciones automatizadas para la detección de contenidos falsos.

En cuanto a las propuestas, se observa un dominio de los trabajos que presentan métodos / sistemas agregados compuestos por diferentes algoritmos (PI1). Gran parte de estos sistemas incluyen sus propias bases de datos construidas *ad hoc*, como consecuencia de la escasez de este tipo de instrumentos en formatos estandarizados que puedan resultar útiles para toda la comunidad investigadora (Saquete et al., 2019). Esta carestía de bases de datos estructuradas se percibe también en nuestro estudio: solo el 11,45% de los trabajos proponen este tipo de herramientas. Sería deseable, por tanto, un mayor desarrollo de estos recursos para alimentar el funcionamiento de los algoritmos, elemento esencial en los modelos lingüísticos, que requieren de este tipo de instrumentos específicos para cada idioma.

Se observa un claro dominio de los modelos algorítmicos lingüísticos (PI1), dada su menor complejidad y, por consiguiente, su mayor trayectoria como objeto investigado. En segundo lugar, destaca la prevalencia de los modelos mixtos, que combinan en diferentes formas los modelos basados en el análisis textual. Menor atención investigadora han suscitado los algoritmos de detección de imágenes, si bien se prevé que el desafío que constituyen las *deepfakes* (vídeos creados mediante inteligencia artificial donde se puede ver a un sujeto expresando opiniones o ejecutando acciones que en realidad nunca llevó a cabo) obligue a un mayor desarrollo de este tipo de soluciones. Es prácticamente nula la investigación en detec-

ción automática del audio *fake*, aspecto que confirma las tesis de García-Marín (2021).

Los algoritmos basados en análisis de redes registran el nivel de precisión más elevado, por encima de los análisis de imágenes y los mixtos (PI2). A pesar de la relativamente extensa investigación en modelos lingüísticos, su nivel de precisión aún resulta mejorable (alcanza el 88,23%, un 1,58% inferior a la media de la muestra analizada). Estos modelos de análisis textual presentan aún varias limitaciones derivadas de la dificultad de detectar las motivaciones psicológicas de los creadores de la desinformación; así como las expresiones de ironía y sátira (Saquete et al., 2019), aspectos difícilmente comprensibles para los modelos algorítmicos. Asimismo, se han observado problemas en la identificación de la desinformación textual generada mediante sistemas de aprendizaje automático porque "los textos falsos creados algorítmicamente son muy similares a los verídicos producidos con estos mismos procedimientos" (Schuster et al., 2020, p. 507).

Las comparaciones estadísticas del nivel de precisión determinan la ausencia de diferencias significativas entre los diferentes modelos analizados, a pesar del distinto grado de complejidad técnica utilizado (PI2). Este hallazgo se alinea de nuevo con el trabajo de Saquete et al. (2019, p. 22), quienes demostraron que los instrumentos más elaborados y sofisticados de *deep learning* "no han demostrado una mejora sustancial, a pesar del coste y los recursos que tales técnicas requieren".

El desembarco de la inteligencia artificial al campo periodístico para la automatización de determinadas labores, como la redacción de noticias, es un hecho que se extiende paulatinamente hacia otras tareas, como la verificación de contenidos y fuentes. Este trabajo constituye una primera aproximación en castellano al análisis de la investigación sobre esta práctica -el *fact-checking* automatizado- centrada fundamentalmente en la tipología de los modelos algorítmicos y su efectividad. Futuros trabajos deberán investigar sobre la mejora y la correcta aplicación de estas soluciones, y su combinación con el talento humano (al que debe apoyar, pero sin ánimo de sustituir), imprescindible para el desarrollo de procesos complejos como es la detección de contenidos falsos, engañosos o sesgados. En todo caso, por su "potencia y velocidad de recopilación y almacenamiento de datos y su capacidad para establecer cartografías precisas de situaciones en curso" (Sadin, 2017, p. 66), estas aplicaciones tecnológicas, situadas en una fase primigenia en cuanto a su desarrollo, ofrecen un potencial prometedor para afrontar los desafíos derivados del fenómeno de la desinformación.

5. Referencias

Aleixandre-Benavent, R., Castelló-Cogollos, L. y Valderrama-Zurián, J.C. (2020). Información y comunicación durante los primeros meses de Covid-19. Infodemia, desinformación y papel de los profesionales de la información. *Profesional de la información, 29*(4), e290408. https://doi.org/10.3145/epi.2020.jul.08.

Amjad, M., Sidorov, G., Zhila, A., Gómez-Adorno, H., Voronkov, I. y Gelbukh, A. (2020). "Bend the truth": Benchmark dataset for fake news detection in Urdu language and its evaluation. *Journal of Intelligent & Fuzzy Systems*, *39*(2), 2457-2469. https://doi.org/10.3233/JIFS-179905.

Antonakaki, D., Fragopoulou, P. y Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, *164*, 114006. https://doi.org/10.1016/j.eswa.2020.114006.

Aparici, R. y García-Marín, D. (2019). *La posverdad. Una cartografía de los medios, las redes y la política*. Gedisa.

Caldevilla-Domínguez, D. y García-García, E. (2020). Profesionales y posverdad: La responsabilidad colectiva como arma contra la falacia digitalizada. *aDResearch, 21*(21), 70-83. https://doi.org/10.7263/adresic-021-04.

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D. y Woźniak, M. (2020). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, 107050. https://doi.org/10.1016/j.asoc.2020.107050.

Dale, R. (2017). NLP in a post-truth world. *Natural Language Engineering, 23*(2), 319-324. https://doi.org/10.1017/S1351324917000018.

Elías, C. (2021). El periodismo como herramienta contra las fake news. En C. Elías y D. Teira (Eds.), *Manual de periodismo y verificación de noticias en la era de las fake news* (pp. 19-57). Editorial UNED.

García-Marín, D. (2021). Las fake news y los periodistas de la generación z. Soluciones post-millennial contra la desinformación. *Vivat Academia. Revista de Comunicación, 154*, 37-63. http://doi.org/10.15178/va.2021.154.e1324.

García-Ortega, A. y García-Avilés, J.A. (2021). Uso del diseño lúdico para combatir la desinformación. *Revista Icono 14, 19*(1), 179-204. https://doi.org/10.7195/ri14.v19i1.1598.

Gereme, F., Zhu, W., Ayall, T. y Alemu, D. (2021). Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information, 12*(1), 20. https://doi.org/10.3390/info12010020.

Goldani, M. H., Momtazi, S. y Safabakhsh, R. (2021). Detecting fake news with capsule neural networks. *Applied Soft Computing, 101*, 106991. https://doi.org/10.1016/j.asoc.2020.106991.

Goldani, M. H., Safabakhsh, R. y Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. *Information Processing & Management, 58*(1), 102418. https://doi.org/10.1016/j.ipm.2020.102418.

Horne, B. y Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1). https://cutt.ly/eQqMXwo.

Huang, Y.F. y Chen, P.H. (2020). Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, *159*, 113584. https://doi.org/10.1016/j.eswa.2020.113584.

Ilias, L. y Roussaki, I. (2021). Detecting malicious activity in Twitter using deep learning techniques. *Applied Soft Computing*, *107*, 107360. https://doi.org/10.1016/j.asoc.2021.107360.

Kasban, H. y Nassar, S. (2020). An efficient approach for forgery detection in digital images using Hilbert–Huang transform. *Applied Soft Computing*, *97*, 106728. https://doi.org/10.1016/j.asoc.2020.106728.

Kaur, S., Kumar, P. y Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), 9049-9069. https://doi.org/10.1007/s00500-019-04436-y

Lease, M. (2018). Fact Checking and Information Retrieval. *DESIRES*, 97-98. https://cutt.ly/rQqMTv3.

López-García, X., Silva-Rodríguez, A., Vizoso-García, A.A. Westlund, O. y Canavilhas, J. (2019). Periodismo móvil: Revisión sistemática de la producción científica. *Comunicar, 2*7, 9-18. https://doi.org/10.3916/C59-2019-01.

Magdy, A. y Wanas, N. (2010). Web-based statistical fact checking of textual documents. *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 103-110. https://cutt.ly/FQqBjDl.

Meneses-Silva, C.V., Silva-Fontes, R. y Colaço-Júnior, M. (2021). Intelligent fake news detection: a systematic mapping. *Journal of applied security research*, *16*(2), 168-189. https://doi.org/10.1080/19361610.2020.1761224.

McIntyre, L. (2018). *Post-truth.* MIT Press.

Mouratidis, D., Nikiforos, M.N. y Kermanidis, K. L. (2021). Deep Learning for Fake News Detection in a Pairwise Textual Input Schema. *Computation*, *9*(2), 20. https://doi.org/10.3390/computation9020020.

Núñez-Mussa, E. (2020). Cómo verificar sin expertos y llegar a las grandes ligas. *Obra Digital, 18*, 85-101. https://doi.org/10.25029/od.2020.236.18.

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 1-6. https://cutt.ly/VQq1bg0.

Oshikawa, R., Qian, J. y Wang, W.Y. (2018). A survey on natural language processing for fake news detection. *arXiv,* 1811.00770. https://cutt.ly/qQq1upc.

Polanco, X. (1997). Infometría e Ingeniería del Conocimiento: Exploración de Datos y Análisis de la Información en vista del Descubrimiento de Conocimientos. En H. Jaramillo y M. Albornoz (Compiladores), *El universo de la medición: La perspectiva de la Ciencia y la Tecnología* (pp. 335-350). COLCIENCIAS, CYTED, RICYT. Tercer Mundo Editores.

Posadas-Durán, J.P., Gómez-Adorno, H., Sidorov, G. y Moreno-Escobar, J. J. (2019). Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4869-4876. https://doi.org/10.3233/JIFS-179034.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. y Stein, B. (2017). A stylometric inquiry into hyperpartisan and fake news. *arXiv,*1702.05638. https://cutt.ly/eQqM8Yr.

Puebla-Martínez, B., Del Campo, E. y Pérez-Cuadrado, P. (2018). Análisis bibliométrico de la revista index.comunicación (2011-2017). Estrategias de posicionamiento inicial. En R. Repiso, J. Guallar y J. M. de Pablos (Eds.), *Revistas científicas de Ciencias de la Información en el abismo* (pp. 39-64). Egregius Ediciones y Universidad de Zaragoza.

Ruiz-Rico, M. (2020). Truth as Literature: Ethics of Journalism and Reality in the Digital Society. *Estudios sobre el mensaje periodístico*, *26*(1), 307-315. https://doi.org/10.5209/esmp.67309.

Sadin, E. (2017). *La humanidad aumentada. La administración digital del mundo.* Caja Negra Editora.

Sahoo, S.R. y Gupta, B.B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983. https://doi.org/10.1016/j.asoc.2020.106983.

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P. y Palomar, M. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert systems with applications*, *141*, 112943. https://doi.org/10.1016/j.eswa.2019.112943.

Schuster, T., Schuster, R., Shah, D. J. y Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, *46*(2), 499-510. https://doi.org/10.1162/COLI a 00380.

Silva, R.M., Santos, R.L., Almeida, T.A. y Pardo, T.A. (2020). Towards automatically filtering fake news in Portuguese. *Expert Systems with Applications*, *146*, 113199. https://doi.org/10.1016/j.eswa.2020.113199.

Thaher, T., Saheb, M., Turabieh, H. y Chantar, H. (2021). Intelligent Detection of False Information in Arabic Tweets Utilizing Hybrid Harris Hawks Based Feature Selection and Machine Learning Models. *Symmetry*, *13*(4). https://doi.org/10.3390/sym13040556.

Zervopoulos, A., Alvanou, A. G., Bezas, K., Papamichail, A., Maragoudakis, M. y Kermanidis, K. (2021). Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests. *Neural Computing and Applications*, 1-14. https://doi.org/10.1007/s00521-021-06230-0.