La IA como herramienta para combatir la desinformación. Planteamiento de un modelo enfocado en los bulos en un contexto electoral

IA as a tool to combat disinformation. Approaching a model focused on hoaxes in an electoral context

ISSN: 1696-019X / e-ISSN: 2386-3978

Mercedes Herrero de la Fuente. Doctora en CC. de la Información (Universidad Complutense de Madrid), Máster en Comunicación Radiofónica (Radio Nacional de España-Universidad Complutense de Madrid) y Máster en Lingüística Aplicada a la Enseñanza de Español como Lengua Extranjera (Universidad Antonio de Nebrija). Investigadora con un sexenio activo y miembro del grupo de investigación INNOMEDIA. Investigadora principal de la Cátedra en Cine, Mujer y Educación, impulsada por EGEDA (Entidad de Gestión de Derechos de los Productores Audiovisuales) y Platino Educa. Integrante del proyecto de investigación I+D+I COM2GENDER, sobre las brechas digitales en la formación universitaria. Ha participado anteriormente en otros proyectos de investigación competitivos con financiación pública, entre ellos COMPENSA, enfocado en la inserción laboral en el sector audiovisual de las personas con discapacidad. Ha sido *research fellow* en Cornell University (EE.UU.), Saldford University (Reino Unido), Radboud Universiteit (Países Bajos) y Univerzita Karlova (República Checa). Publica artículos en revistas de alta indexación centrados en: aplicación de la tecnología al discurso informativo, nuevos perfiles profesionales y participación de la mujer en el sector audiovisual. En la actualidad es Coordinadora del Doctorado en Innovación en Comunicación Digital y Medios y profesora acreditada en la Universidad Nebrija. Ha sido durante quince años productora de informativos en Telemadrid.
Universidad Nebrija, España
mherrero@nebrija.es
ORCID: 0000-0002-5361-9056

Celia Sancho Belinchón. Doctora en Comunicación Audiovisual, Publicidad y Relaciones Públicas. Actualmente dirige el Máster Universitario en Periodismo digital y de datos en la Universidad Nebrija, además imparte docencia en el grado de Periodismo y en el Máster Universitario en Comunicación Política y Gestión de Crisis y Emergencias. Sus líneas de investigación abarcan el periodismo, la comunicación digital, redes sociales, estrategia de medios sociales, *fact-checking* y publicidad.
Universidad Nebrija, España
csanchobe@nebrija.es
ORCID: 0000-0001-5979-1853

Cómo citar este artículo:

Herrero de la Fuente, M.; Sancho Belinchón, C. y Sedeño López, J. (2025). La IA como herramienta para combatir la desinformación. Planteamiento de un modelo enfocado en los bulos en un contexto electoral. *Doxa Comunicación*, 41, pp. 511-533.

https://doi.org/10.31921/doxacom.n41a2840

Este contenido se publica bajo licencia Creative Commons Reconocimiento - Licencia no comercial. Licencia internacional CC BY-NC 4.0

ISSN: 1696-019X / e-ISSN: 2386-3978

Jorge Sedeño López. Doctor en Ingeniería Informática, Máster en Gestión de las Tecnologías de la Información y las Comunicaciones e Ingeniero en Informática por la Universidad de Sevilla. Cuenta con una dilatada experiencia profesional de más de veinte años en diferentes Administraciones Públicas y pertenece al Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado y al Cuerpo de Expertos en Tecnología de la Información del Banco de España. Sus líneas de trabajo están orientadas a la transformación digital mediante las metodologías ágiles, el Gobierno Electrónico, el Gobierno del Dato y la Inteligencia Artificial. Pertenece al grupo de investigación ES3 (Engineering and Science in Software System) y es miembro del Proyecto EQUAVEL PID2022-137646OB-C31 financiado por MICIU/AEI /10.13039/501100011033 y por FEDER, UE. Ejerce la docencia en el Master de Periodismo Digital y de Datos en la Universidad de Nebrija y en otras instituciones públicas y privadas.
Universidad de Sevilla, España ROR
jorgesedeno@us.es
ORCID: 0000-0002-5368-5547

Recibido: 20/11/2024 - Aceptado: 14/06/2025 - En edición: 19/06/2025 - Publicado: 01/07/2025

Received: 20/11/2024 - Accepted: 14/06/2025 - Early access: 19/06/2025 - Published: 01/07/2025

La inteligencia artificial (IA) ha contribuido a la desinformación por su capacidad para generar contenidos falsos. Pero el potencial de esta tecnología puede también enfocarse en diseñar un prototipo de herramienta que detecte los bulos, en concreto aquellos amplificados en redes sociales y en contextos electorales. Este artículo analiza los principales patrones seguidos por las noticias falsas lanzadas en X durante las últimas elecciones catalanas (12 mayo 2024), siguiendo criterios como la temática, el formato, el origen o su difusión, entre otros. Con la información obtenida se elabora de forma preliminar un recurso de IA con capacidad de reconocer tales contenidos. Partimos de estos resultados concretos: el tema más recurrente es la inmigración, predomina el formato texto más fotografía, en la mayoría de los casos procede de perfiles registrados como un ciudadano cualquiera y los medios convencionales no participan, en general, en su propagación. Sobre estas pautas planteamos las principales características de un sistema IA que combina patrones de difusión con análisis de texto, imágenes y sentimiento, que junto con la verificación en tiempo real de hechos nos permita filtrar con un grado suficiente de sensibilidad (proporción de bulos correctamente identificados) y especificidad (proporción de contenidos veraces erróneamente clasificados como bulos).

Palabras clave:

Inteligencia artificial; desinformación; verificación; elecciones catalanas; algoritmo.

Abstract:

Artificial intelligence (AI) has contributed to disinformation through its ability to generate false content. But the potential of this technology can also be focused on designing a prototype tool that detects hoaxes, particularly those amplified in social networks and in electoral contexts and moments of political relevance. This article analyses the main patterns followed by the fake news launched on X during the last Catalan elections (12 May 2024), following criteria such as subject matter, format, origin and dissemination, among others. With the information obtained, an AI resource with the capacity to recognise such content is preliminarily developed. We start from these specific results: the most recurrent topic is immigration, the text plus photograph format predominates, in most cases it comes from profiles registered as any citizen, and the conventional media do not generally participate in its propagation. Based on these guidelines, we propose the main characteristics of an AI system that combines dissemination patterns with analysis of text, images and sentiment, which, together with real-time verification of facts, allows us to filter with a sufficient degree of sensitivity (proportion of hoaxes correctly identified) and specificity (proportion of truthful content erroneously classified as hoaxes).

Keywords:

Artificial Intelligence; disinformation; verification; Catalan elections; algorithm.

1. Introducción

La inteligencia artificial (IA) está presente en múltiples ámbitos de nuestra sociedad, especialmente en el sector periodístico (Lopezosa-García et al., 2024). Los primeros ejemplos de su uso en las redacciones se remontan a 2014, cuando la agencia Associated Press comenzó a utilizarla para resúmenes deportivos y reportajes sobre negocios (Badgamia, 2023). Sin embargo, es en la actualidad cuando el debate sobre la IA generativa ha cobrado más fuerza, por su potencial para agravar el fenómeno de la desinformación. Este trabajo parte de esta problemática, pero se centra en las posibilidades de la IA como herramienta que ayude a la detección de noticias falsas.

Según la UNESCO (Organización de Naciones Unidas para la Educación, la Ciencia y la Cultura) (2021), la IA es la "simulación de procesos de inteligencia humana por parte de máquinas". Tales desarrollos incluyen el aprendizaje, el razonamiento y la autocorrección. En relación con esta definición, Blanco-Marañón (2023) incide en el término "simulación", afirmando que imitar no significa ser igual. Criado-Grande entiende que lo anterior hace referencia a la posibilidad de que las máquinas alcancen "algún tipo de racionalidad mediante la percepción del ambiente con el que interaccionan" (2021, p.351), usando sensores, obteniendo y procesando datos, razonando sobre los mismos y adoptando decisiones.

En el entorno de la comunicación, uno de los cambios acelerados por la IA es el carácter cada vez más líquido de la información. Así, "veremos textos que se convierten en imágenes, audio o vídeo, lo que supondrá la alteración de los modelos y procesos de producción, distribución y monetización como nunca se había experimentado antes" (Cerezo-Guilarranz, 2024, p.49). El potencial de la IA se presenta como un arma de doble filo que puede funcionar como una herramienta eficiente para el periodismo, pero también convertirse en una amenaza por su competencia en la generación de contenidos falsos en cualquier formato.

Desde el punto de vista técnico, será necesario identificar las fuentes de datos, que, en el caso de la desinformación, estarán relacionadas con la web y las redes sociales. También saber cómo se almacenará y accederá a esa información. Según el estudio de IDC (International Data Corporation), la "data esfera" llegará a 175 ZB[1] en 2025 (Reinsel et al., 2018, p.3) lo que representa un gran desafío. Igualmente es preciso conocer de qué manera se va a tratar esta información, campo donde entra en juego la ciencia de datos (inteligencia artificial y aprendizaje de las máquinas) para desarrollar indicadores del grado de falsedad/veracidad de la información. Por último, se ha de definir la forma de extraer y presentar el valor obtenido del tratamiento anterior y su grado de calidad.

1.1. Desinformación e IA

La CE (Comisión Europea) define la desinformación (*disinformation*) como cualquier forma de "información falsa, inexacta o engañosa, diseñada, presentada o promovida para causar intencionadamente un daño público o para obtener un beneficio" (2018, p.3). Se incide pues en la intencionalidad de causar un perjuicio o de obtener una ganancia, aunque sea de forma no ética. Para Alandete-Ballester (2019), las noticias falsas no tienen por qué ser una mentira absoluta; suelen tener alguna vinculación con lo que está pasando, pero se caracterizan por deformar la realidad persiguiendo el sensacionalismo.

1 1 ZB (Zettabyte) = 1.000.000.000.000 GB (Gigabyte).

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

El fenómeno de la desinformación es una seria preocupación para los países democráticos (Rodríguez-Martelo et al., 2023). A través de los bulos se intenta manipular a la ciudadanía y socavar las principales instituciones políticas (Arrieta-Castillo y Rubio-Jordán, 2023). El Libro Blanco contra la Desinformación publicado por el Gobierno de España en 2022 alerta contra la amenaza que supone para la estabilidad política y la seguridad nacional, "debido a su potencial para corromper el debate público, erosionar la confianza en las instituciones, manipular a la opinión pública y condicionar la política exterior" (2022, p.9).

La desinformación concierne cada vez más a la opinión pública, según muestran distintos informes nacionales e internacionales. El Instituto Reuters afirma que esta inquietud ha crecido en 2023 dos puntos respecto al año anterior y que el 56% de los consultados temen no distinguir entre lo que es verdadero y lo que es falso, cuando consumen noticias en Internet (Newman, 2024, p.17).

La IA posee herramientas sofisticadas que pueden usarse para amplificar este indeseable fenómeno. Las más conocidas en los últimos años han sido la generación automática de texto y los *bots*. Ambas han sido aplicadas en redes sociales para fabricar de forma masiva textos engañosos y difundirlos de manera incansable a través de perfiles falsos, amplificando su alcance.

El elemento más novedoso en la actualidad es el *deepfake* o pieza de vídeo y audio en la que las imágenes y el sonido (normalmente ambos) han sido manipulados (Herrero-De-La-Fuente y Ríos-Calvo, 2022). Como señala Deeptrace (2019), las primeras creaciones de este tipo surgen en noviembre de 2017, al crearse en Reddit un foro con el mismo nombre centrado en el uso de programas de *deep learning* para editar vídeos pornográficos. Desde entonces, las herramientas de IA disponibles para este tipo de montajes no han cesado de crecer, siendo algunas de fácil acceso y manejo. Podemos decir que prueba de ello es que la circulación de *deepfakes* ha crecido un 550% entre 2019 y 2023, según la organización para la seguridad en línea Home Security Heroes (2024). Los *deepfakes* representan una ruptura en la confianza que la sociedad tiene en la imagen (Jacobsen y Simpson, 2023), por ello constituyen un cambio de paradigma, donde deja de tener sentido aquello de "ver para creer"[2].

Como hemos apuntado, las redes sociales son un agente esencial en la diseminación de desinformación. En los últimos años los medios de comunicación han perdido el monopolio en la distribución de noticias, de forma que los contenidos informativos han proliferado fuera de los circuitos mediáticos y son propagados por millones de cuentas creadas por individuos, grupos políticos, empresas o cualquier organización (González-Quintero y Cardona-Restrepo, 2023), que muestra o no su verdadera identidad. En nuestro país, el número de personas que utilizan a diario Internet (87%) es ya mayor que el que ve la televisión (81%) (AIMC, 2024, p.38 y p.64). El acceso a la información de actualidad figuraba precisamente entre los principales usos de Internet en diciembre de 2023 (así lo declaran el 60% de los preguntados), al tiempo que un 70% la emplea para navegar por las redes sociales y un 97% para la mensajería instantánea (AIMC, 2024, p.67). Se debe añadir que esta última, debido a su carácter cerrado, constituye uno de los principales canales de difusión de *fake news* (Díez-Garrido *et al.*, 2021).

Además, las noticias elaboradas en el ámbito periodístico, es decir, por periodistas o expertos, pierden peso en el conjunto de contenidos supuestamente informativos que se consumen en redes sociales, especialmente entre los más jóvenes (Newman, 2024, p.11). Según el último informe del Instituto Reuters, "mientras que los periodistas convencionales suelen liderar las conversaciones en torno a las noticias en Twitter y Facebook, apenas llaman la atención en redes más nuevas como Instagram,

2 Uno de los más sonados en los últimos años fue el vídeo manipulado del presidente ucraniano, Volodímir Zelenski, declarando la rendición de su país frente a Rusia en los primeros días de la guerra iniciada en febrero de 2022. https://cutt.ly/0w2geeXl

Snapchat y TikTok" (Newman, 2024, p.13), donde los *influencers* acaparan el mayor protagonismo. Por tanto, este dato marca una tendencia también registrada por el Edelman Trust Barometer, que detecta en su último estudio global un aumento de la desconfianza en los medios en quince de los veintiocho países consultados. Italia, Alemania y Brasil son los tres donde este descrédito mediático es mayor, situándose España sólo cuatro puestos después (2024, p.43).

1.2. La IA como instrumento para combatir la desinformación

Peña-Fernández *et al*. (2023) señalan entre las principales aplicaciones de la inteligencia artificial el desarrollo de herramientas para detectar la desinformación. Según explica García-Marín, "la IA permite determinar la credibilidad de las fuentes informativas a partir de su análisis reputacional, a la vez que ofrece una poderosa respuesta para identificar perfiles falsos en redes sociales" (2021, p.53). Igualmente es capaz de detectar contenidos desinformativos mediante el uso de la lingüística computacional (con modelos semánticos y sintácticos) y de métodos no lingüísticos para descubrir manipulaciones de imágenes (fotografías o vídeos).

Junto al *big data*, la IA puede ser un instrumento para desenmascarar contenidos falsos, tal como señalan algunas investigaciones recientes (Moreno-Espinosa et al., 2024; García-Marín, 2021; Flores-Vivar, 2019). Entre los dispositivos más utilizados para localizar *fake news* destacan distintos tipos de *bots*, desarrollados en muchos casos en colaboración entre universidades, empresas y medios. Se basan en aspectos como algoritmos con capacidad de adaptación, que examinan fuentes y patrones de difusión, principalmente. Algunos ejemplos son Fact Machine (del verificador brasileño Aos Fatos), TruthBuzz (impulsada por el International Center for Journalists) o Les Décodeurs (del diario *Le Monde*). Recientemente se han desarrollado numerosas herramientas de IA fundamentadas en el *machine-learning*, que trabajan con patrones lingüísticos mediante clasificadores de aprendizaje automático, indicando la veracidad de una noticia en función de distintas variables (Luengo-Cruz y García-Marín, 2020). Mencionamos aquí, por citar sólo una de ellas, Fakebox, que discrimina entre artículos escritos de forma similar a las noticias reales y textos que no siguen esas pautas, adjudicando una puntuación (Telefónica Tech, 2018). Otros sistemas novedosos son ClaimBuster (https://idir.uta.edu/claimbuster/) y Full Fact (https://fullfact.org/) este último especialmente diseñado para contenido político con sistemas en tiempo real y bases de datos de hechos verificados (aunque no en castellano). En el presente artículo, en la exposición de nuestros resultados, profundizamos más sobre este tipo de recursos.

Las plataformas de redes sociales lideran numerosos proyectos que desarrollan sistemas de IA para eliminar automáticamente contenido malicioso a través del análisis basado en texto; mencionamos entre ellos, Facterbot o Projeto Lupe (Flores, 2019). Sin embargo, muchas publicaciones son fotos, vídeos o audios, para los que estos métodos de chequeo no se encuentran aún tan desarrollados (Moreno-Espinosa et al., 2024). Esto implica que la identificación de la desinformación es un desafío importante en el campo del aprendizaje automático (*machine-learning)* y la inteligencia artificial. Se puede decir que no hay un sólo "mejor" algoritmo para esto, ya que el enfoque adecuado depende de varios factores, como el tipo de datos y formatos disponibles, la complejidad del problema y las características específicas de las noticias falsas que se intentan detectar, siendo necesario, además, un abordaje multidisciplinar (Ruffo et al., 2023).

Desde la Unión Europea se han impulsado iniciativas para combatir la desinformación desde diferentes enfoques. Ya en 2018 se establecieron una serie de ámbitos de actuación en: investigación específica sobre este fenómeno en las diferentes áreas

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

implicadas (existe desde 2018 el proyecto Fandango), formación de redes independientes de verificadores (nace en 2019 FactCheckEU) y promoción de la alfabetización mediática, entre otros. Las instituciones comunitarias apuestan por enfrentar las noticias falsas con una estrategia transversal, que va más allá de la creación de herramientas tecnológicas para la detección de bulos. Tales recursos han de apoyarse en los profesionales de la información y en todos los ciudadanos. "La clave es contar con una ciudadanía que entienda la importancia de obtener información de calidad de fuentes solventes, que sea capaz de identificar los contenidos potencialmente falsos y, en suma, que valore la verdad" (Sádaba y Salaverría, 2023, p.27). A este respecto, el nuevo Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de IA menciona los posibles efectos negativos "sobre el marco democrático, el discurso cívico y los procesos electorales" (BOE, 12 julio 2024).

2. Método

El presente estudio tiene como principal objetivo extraer las principales características de los bulos difundidos en los últimos comicios autonómicos en Cataluña (12 mayo 2024) y extrapolar dichas reglas para diseñar un prototipo de herramienta de IA con la funcionalidad de detectar bulos relacionados con esta tipología de contexto electoral.

Para ello desarrollamos nuestra investigación en tres fases. La primera, de carácter descriptivo, se centra en la consulta de informes sectoriales e investigaciones previas. Se lleva a cabo una revisión de la literatura relacionada con nuestro objeto de estudio y los datos más actuales al respecto y se establecen así las bases para una reflexión dirigida a elaborar el marco conceptual sobre el que se asienta nuestro trabajo.

La segunda consiste en un análisis de contenido de los bulos más populares en la red X (antes Twitter) durante el periodo de precampaña (iniciado el 2 de abril), campaña (del 3 al 10 de mayo) y jornada de las elecciones autonómicas catalanas (12 de mayo de 2024). La selección de los bulos (catorce) se ha realizado a partir de un artículo del blog Maldito Bulo (Maldita, 2024), que recoge los que mayor repercusión han tenido en redes sociales (en concreto en X). Se estima que esta muestra representa un conjunto mayor, puesto que reúne los más significativos. Se otorga crédito a la recopilación llevada a cabo por Maldita por ser una fundación dedicada a la verificación de reconocido prestigio en nuestro país[3]. No obstante, proponemos un muestreo intencional, que sigue un criterio de facilidad, ya que se trata de una muestra accesible para llegar a la cual Maldita ha aplicado ya unos filtros. Se considera que es un punto de partida para esta fase inicial de desarrollo de un prototipo de herramienta de IA, tal como señalamos al explicar la tercera fase de la investigación.

Se debe añadir que todos los contenidos analizados están escritos en español a excepción de los publicados por el canal de televisión TV3, que son en catalán, aunque dicho medio sí retuitea textos en español.

3 Maldita nace como medio nativo digital sin ánimo de lucro en 2018 y desde sus inicios colabora con la Comisión Europea en varias iniciativas para luchar contra la desinformación. Es pionera en la labor de *fact-checking* en nuestro país y cuenta con reconocimiento internacional, siendo miembro de la International Fact-Checking Network.

El patrón de análisis que se ha seguido para analizar los catorce bulos destacados ha sido el siguiente:

Tabla 1. Variables de análisis de bulos

Variable	Descripción
Fecha	Día de publicación del bulo
Temática	Área temática a la que se refiere el bulo*
Partido político	Formación política objeto del bulo
Formato	Texto; texto y fotografía; fotografía editada; texto y vídeo; vídeo; texto y audio
Redacción del bulo	Modo en el que se aborda la información (directo, indirecto, explicativo, sensacionalista)
Tipo de perfil que origina el bulo	Portavoz político (cargo público, cargo partido u organización afín); medio de comunicación; ciudadano desconocido
Difusión del bulo por los medios de comunicación	Sí/No Se consideran medios digitales en la red social X. Se analizan medios tradicionales, como: periódicos (*El País, El Mundo, ABC, La Razón, El Diario, Público*), radios (*SER, COPE, Onda Cero, RNE*) y televisiones (*La 1, AT3, Telecinco, La Sexta y TV*3)

***Optamos por dejar esta variable abierta, sin definir categorías que puedan predeterminarla**

Fuente: elaboración propia

La tercera fase, de carácter exploratorio, propone, una vez analizados los bulos seleccionados e identificados los patrones comunes, el diseño de un prototipo de herramienta de IA que lleve incorporados los mecanismos para identificar dichos patrones, con el fin de detectar los mismos en posteriores noticias y determinar con una sensibilidad y especificidad suficientes si se trata o no de un bulo.

La sensibilidad es la tasa de verdaderos positivos y la especificidad es la tasa de falsos positivos, de manera que para un modelo de clasificación binaria como el que se propone (información falsa, información cierta) cuanto mayor sea la sensibilidad (más aciertos) y menor sea la especificad (menos fallos), más se acercará a un clasificador perfecto.

Para cada uno de estos patrones se propondrán una serie de algoritmos que se ajustarán con hiperparámetros. Un hiperparámetro es una configuración ajustable externamente, que no es aprendida a partir de los datos, sino que se establece antes del proceso de entrenamiento del modelo, es decir, es un conjunto de parámetros externos al propio algoritmo, que puede mejorar o adaptar el rendimiento de este (Simanjuntak et al., 2024) respecto al problema propuesto.

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

Así mismo, es especialmente relevante, dado que hay muchas combinaciones y ajustes, la elección de métricas de desempeño relacionadas con la sensibilidad y la especificidad en un modelo de clasificación binaria (información falsa-bulo, información cierta-no bulo).

Por último, para abordar el problema de la clasificación de bulos y entender por qué se toman ciertas decisiones, los modelos explicables (o interpretables) son cruciales. Estos modelos no sólo deben ser precisos, sino también proporcionar una justificación clara para sus predicciones, por lo tanto, el sistema deberá tener la capacidad de explicar la decisión tomada (Hashmi et al., 2024).

3. Resultados

3.1. Principales bulos difundidos en X con relación a las elecciones autonómicas de Cataluña

Iniciamos la exposición de este primer estadio de los resultados mostrando una ficha con los datos identificativos ligados estrechamente con el contenido de los catorce principales bulos detectados, que recoge la Tabla 2.

Tabla 2. Ficha identificativa principales bulos en las elecciones de Cataluña (fecha, tema, partido)

Denominación	Descripción del bulo	Fecha	Temática	Partido político[4]
Casa Tarradellas	Declaraciones falsas del fundador de la empresa explicando que sólo contrata a catalanes	02/04/2024	Cataluña/ Marca	Ninguno
Jordi Évole ficha por ERC	Noticia falsa sobre el fichaje del periodista por ERC	09/04/2024	Fichajes partidos	ERC/ Gobierno
Abuelo Puigdemont	Fotografía del falangista Gregorio Martín Mariscal mal atribuida	17/04/2024	Candidatos	Junts+ Puigdemont
Vacunación COVID Illa	Declaraciones tergiversadas sobre su vacunación	22/04/2024	Candidatos/ Salud	PSC
Epidemia tiña	Alerta sanitaria falsa por epidemia de tiña	24/04/2024	Salud	Ninguno
Carteles “emirato islámico”	Imagen falsa de carteles de bienvenida situados en la entrada a distintas localidades	25/04/2024	Inmigración	Frente Obrero

4 ERC: Esquerra Republicana de Catalunya; PSC: Partido Socialista de Catalunya, perteneciente al PSOE: Partido Socialista Obrero Español; Junts: Junts per Catalunya.

Acto bandera Illa	Imagen real sacada de contexto de una marcha por la unidad de España con diferentes partidos políticos presentes	26/04/2024	Candidatos	PSC
Ayudas sociales a marroquíes	Noticia falsa que indica que las familias de marroquíes sólo viven de ayudas en Cataluña	26/04/2024	Subvenciones/ Inmigración	ERC /Gobierno
Renta Garantizada	Datos falsos sobre las ayudas a familias marroquíes del antiguo PIRMI	03/05/2024	Subvenciones/ Inmigración	ERC /Gobierno
Implantación árabe en colegios	Falsa subvención de la Generalitat para implantar el idioma	03/05/2024	Educación/ Inmigración	ERC /Gobierno
Compra votos asesor PSOE	Supuesta compra de votos a través de WhatsApp del asesor	10/05/2024	Corrupción	PSOE y PSC
Piscinas privadas expropiadas	Noticia falsa de expropiación de piscinas privadas a consecuencia de la sequía	10/05/2024	Sequía	ERC /Gobierno
Ayuda PIRMI (Renta Mínima de Inserción)	Ayudas sociales a familias de cualquier nacionalidad, que ya no existe en Cataluña	10/05/2024	Subvenciones/ Educación / Inmigración	ERC /Gobierno
Sabotaje Rodalies[5]	Supuesto robo de cables de cobre de los Rodalies dejan paralizado el servicio	12/05/2024	Transporte / Corrupción	PSOE

Fuente: elaboración propia

Como se puede observar en la tabla, encontramos noticias falsas repartidas entre el 2 de abril de 2024 y el 12 de mayo de 2024, aunque no de manera equitativa, puesto que no se publican todos los días. Únicamente se han encontrado los bulos señalados en la Tabla 2, siendo la fecha con más noticias falsas el 10 de abril. El periodo abarca desde la precampaña hasta la jornada electoral, ya que la tónica de confrontación de los últimos años ha ampliado en el tiempo la diatriba partidista. Llama la atención que el clima de crispación se intensifique el mismo día de los comicios.

Comprobamos que las áreas temáticas más recurrentes son las relativas a los temas sociales y la cuestión más presente es la inmigración, que está detrás de cinco contenidos relacionados con subvenciones/ayudas económicas y educación. Dentro de la temática social también se alude a la salud, pero sólo en el caso de la falsa epidemia de tiña se trata como tal, sin utilizarse como excusa para otras cuestiones. Igualmente encontramos alusiones a candidatos y fichajes de las distintas formaciones (cuatro). Así mismo, existen bulos que aparentemente recurren a cuestiones fuera de la política, como una marca (Casa Tarradellas) o el problema de la sequía, para lanzar contenidos relativos al nacionalismo catalán o a políticas económicas en contra de la

5 Se denomina Rodalies al servicio público de trenes de cercanías y regionales media distancia de Cataluña. Fue traspasados por el Ministerio de Fomento a la autonomía catalana en 2010 y 2011.

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

propiedad privada, todo ello con el fin de descreditar al gobierno central. Con el mismo propósito, se habla de compra de votos por WhatsApp o de incidencias en el transporte de cercanías durante la jornada electoral.

Como se refleja en la Tabla 2, únicamente se han encontrado dos bulos que no corresponden de manera directa a un partido político en concreto. De los doce restantes: seis aluden al partido ERC (que gobierna Cataluña en el momento de realizar esta investigación), tres al PSC (uno de ellos compartido con el PSOE), uno al PSOE, uno a Junts y otro al Frente Obrero. Todo ello parece obedecer también a una estrategia de desgaste de las instituciones, al intentar erosionar a los partidos que lideran tanto el ejecutivo autonómico como el central.

Dentro de la muestra contemplada el formato más utilizado es la combinación de texto y fotografía con un total de diez casos (la imagen no presenta ningún tipo de edición o de texto superpuesto). Hay que aclarar que, en dos más el cuerpo del tuit consta únicamente de una fotografía (sin texto) y esta aparece con un texto incrustado.

Encontramos dos formatos más en los catorce bulos analizados, que son: audio acompañado de texto, recogiendo el testimonio sonoro de un inmigrante de origen magrebí, que recibía ayudas sociales (PIRMI) de manera indebida; y vídeo combinado con texto, en el tuit sobre la supuesta negativa de Salvador Illa, candidato socialista, a vacunarse de COVID-19.

Se debe señalar que existen algunos formatos de los que no se han obtenido resultados, como son sólo texto o sólo vídeo. En cualquier caso, ninguna de las publicaciones parece asociada a un usuario con una alta competencia en habilidades digitales.

Por norma general, los textos suelen ser sensacionalistas, utilizando mayúsculas, exceso de signos de puntuación y faltas de ortografía en algunos casos. En la mayoría, dichos textos pueden ayudar a captar más la atención del espectador, ya que utilizan colores vivos como el rojo. Por lo tanto, se puede decir que esta estética sencilla y basada en lo visual consigue un mensaje muy efectivo y es fácilmente asimilado por la audiencia, lo que contribuye al engaño. Así mismo, no se han encontrado bulos redactados de modo indirecto, sino que todos están escritos en estilo directo, basados en argumentos simplistas para captar la atención del espectador. Tampoco incluyen explicaciones, ya que la estrategia es la simplificación y la exposición de argumentos maximalistas.

Si atendemos a los perfiles de X que han publicado los bulos analizados, observamos una coincidencia casi en su totalidad, ya que la gran mayoría (once) han sido originados en dicha red social por usuarios desconocidos en la esfera pública. Se salen de este patrón tres casos: el primero, creado por el líder de Frente Obrero; el segundo, generado por portavoces políticos del partido VOX; y el tercero diseñado por portavoces políticos del Partido Popular. En los tres casos los miembros de dichos partidos políticos no estaban en posesión de ningún cargo público en el momento de la difusión de estos bulos.

Dentro de los tres bulos lanzados por cargos políticos de los partidos indicados, el primero, originado por Roberto Vaquero, muestra unos carteles de bienvenida en las entradas por carretera de la Comunidad Autónoma de Cataluña con el enunciado "Emirato Islámico de Cataluña". Nace del perfil en X de este líder del Frente Obrero (partido ultraderechista). A su vez, ciudadanos que pueden identificarse como afines a dicha información lo difunden en sus cuentas personales.

El segundo se enfoca en el boicot a los trenes de cercanías catalanes por parte del PSOE, para evitar que los ciudadanos se desplacen y ejerzan su derecho al voto. Dirigentes del PP lanzan acusaciones en este sentido que se ven retuiteadas y amplificadas por miembros de ERC, Junts y Frente Obrero.

El último bulo hace referencia a la supuesta inversión que realiza el Gobierno autonómico de Cataluña para implantar el estudio del idioma árabe en los colegios. Se han encontrado publicaciones con esta noticia falsa en perfiles de X pertenecientes a portavoces de VOX y a seguidores de esta formación (no identificados como afiliados).

En lo relativo a las publicaciones de las noticias falsas analizadas en otros medios de comunicación, en términos generales, los medios no se han hecho eco de las mismas en sus perfiles oficiales de X; pero encontramos dos excepciones a esta pauta.

La primera, en la noticia falsa sobre Salvador Illa y el supuesto hecho de no haber acudido a un centro de salud para vacunarse contra el COVID-19. En este caso, ha sido un medio hondureño televisivo, "Girasol TV", el que la incluye en su perfil de X.

La segunda tiene que ver con el presunto sabotaje a los Rodalíes. Los medios tradicionales (indicados en la pauta de análisis), así como la televisión autonómica de Cataluña (TV3) recurren al *clickbait*, publicando en la red X titulares con los términos "supuesto sabotaje" o "sabotaje electoral". Después se refieren al robo de cobre sufrido realmente por la red ferroviaria de cercanías, pero en ningún caso reproducen el núcleo del bulo, es decir, que el partido socialista está detrás de los problemas en los desplazamientos por tren durante la jornada electoral.

Con el objetivo de cuantificar las diferentes variables analizadas mostramos un resumen a continuación:

Tabla 3. Cuantificación de resultados en variables de análisis (no relacionadas directamente con: tema, fecha, partido)

Variable	Descripción	Número de bulos
Formato	Texto	0
	Texto y fotografía	10
	Fotografía editada	2
	Texto y vídeo	1
	Texto y audio	1
Redacción del bulo	Directo	14
	Indirecto	0
	Explicativo	0
	Sensacionalista	14*

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

Tipo de perfil que origina el bulo	Portavoz político	3**
	Medio comunicación	0
	Usuario desconocido	11
Difusión del bulo por los medios	Sí difunden	12
	No difunden	2

*** Todos los bulos pertenecen a ambas categorías: directo y sensacionalista.**

**** Ninguno de estos dirigentes políticos ostenta un cargo público.**

Fuente: elaboración propia

Para aportar mayor claridad sobre la repercusión de los bulos analizados en la muestra, se pueden consultar los datos al respecto en la siguiente tabla:

Tabla 4. Repercusión de bulos en X

Denominación	Perfil que origina el bulo	Difusión del bulo por los medios de comunicación	Número de reposteados	Comentarios
Casa Tarradellas	Usuarios desconocidos	n/p*	134.000	875
Jordi Évole ficha por ERC	Usuarios desconocidos	n/p	42	3
Abuelo Puigdemont	Usuarios desconocidos	N.A.	84.000	28
Vacunación COVID Illa	Usuarios desconocidos	Girasol TV	145.000	1.714
Epidemia tiña	Usuarios desconocidos	n/p	458	54
Carteles “emirato islámico”	Roberto Vaquero (Frente Nacional)	n/p	10	0
Acto bandera Illa	Usuarios desconocidos	n/p	2.600	369
Ayudas sociales a marroquíes	Usuarios desconocidos	n/p	2.698	8
Renta Garantizada	Usuarios desconocidos	n/p	1.293	83
Implantación árabe en colegios	Portavoces políticos de VOX	n/p	1.258	25

Compra votos asesor PSOE	Usuarios desconocidos	n/p	1.000	256
Piscinas privadas expropiadas	Usuarios desconocidos	n/p	1.455	6
Ayuda PIRMI (Renta Mínima de Inserción)	Usuarios desconocidos	n/p	987	18
Sabotaje Rodalies	Portavoces políticos del Partido Popular	*TV3, El País, ABC, La 1, AT3, Telecinco, La Sexta y TV3*	162.000	493

*** No procede**

Fuente: elaboración propia

Como se observa en la tabla hay tres publicaciones que destacan claramente en su expansión sobre el resto y que aluden a temas muy relacionados con Cataluña, intentando desacreditar a un empresario y un político de esta comunidad y avivando así el sentimiento anticatalán. La de mayor alcance ataca a los trenes de cercanías catalanes, pero en este caso el objetivo es el gobierno de España y el generador del bulo el Partido Popular.

3.2. Algoritmos y features para la elaboración de una herramienta de detección de bulos en elecciones autonómicas en España

Los patrones detectados en el análisis de la fase anterior sirven como base para la tercera etapa, que se focaliza en el diseño de un recurso de IA, concatenando diferentes algoritmos, para que sea capaz de identificar esta tipología de bulos escogidos. Dado que la muestra adolece de una limitación numérica, la herramienta propuesta se considera en un estadio inicial de desarrollo.

De esta manera y partiendo de los bulos analizados *a posteriori* en el contexto de unas elecciones autonómicas en España, se esbozará propondrá el diseño de un sistema de detección basado en las pautas identificadas, para lo que será necesario utilizar paradigmas de la IA**,** mediante algoritmos *y features,* (una característica o propiedad individual de los datos de entrada que ayuda al modelo a reconocer patrones y hacer predicciones) que sean capaces de:

– Analizar la difusión de medios, para detectar el origen de la información que vamos a procesar (anónimos o de portavoces políticos sin cargo público).
– Detectar formatos y estilos en la información y analizar el sentimiento (información multiformato con estilo directo y sensacionalista, que busca provocar polarización).
– Verificar hechos (*fact-checking*).

Desgranaremos de manera más detallada estos conceptos a continuación.

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

3.2.1. Difusión de medios

Dentro del análisis de la difusión de medios se propone la utilización del algoritmo PageRank basado en el concepto de centralidad de eigenvector[6], para identificar cuentas influyentes y patrones de diseminación de bulos.

Como *feature* principal optamos por el análisis de grafos, para modelar las relaciones entre esas cuentas y los patrones de difusión, dado que la generación de los bulos se ha realizado mediante la red X y buscamos cuentas anónimas o de portavoces políticos que no ostentan ningún cargo público.

3.2.2. Formatos y estilos

Dado que el patrón de formato identificado en los bulos de la fase dos combina texto y otro elemento, como fotografías (editadas o no), vídeos o audio (que se transformará posteriormente a texto). Se proponen los siguientes algoritmos:

– Para el texto se utilizará un modelo de transformador, es decir, un tipo de arquitectura de red neuronal para el procesamiento del lenguaje natural, por su capacidad para manejar secuencias de texto de manera eficiente y establecer relaciones complejas entre las palabras de una oración. Tanto BERT (Bidirectional Encoder Representations from Transformers) como RoBERTa (A Robustly Optimized BERT Pretraining Approach) pueden ser ajustados para la detección de desinformación (Zhang et al., 2024). Se recurrirá además a modelos preentrenados (HugginFace, FakeBERT, o FakeNewsBERT) junto con los *embeddings*[7] incorporados del análisis de bulos para mejorar la detección de estos (Yang et al., 2018). De esta manera podemos tener modelos híbridos enfocados al contexto de las elecciones autonómicas españolas. Es lo que se conoce como una arquitectura RAG[8].

– Para las imágenes y vídeos se usará EfficientNet, una familia de modelos de redes neuronales convolucionales (CNN), que han demostrado ser muy efectivos en la clasificación de estos contenidos, con capacidad de capturar detalles finos, que son críticos para la detección de manipulaciones (fotografía editada) y que pueden integrarse fácilmente con modelos transformadores, formando un sistema robusto de detección multimodal. Así mismo, el modelo EfficientNet B4 ha demostrado una precisión de detección superior al 92%, de rostros manipulados en vídeos, ya que analiza diversos parámetros, como las expresiones faciales y las irregularidades en las imágenes, para diferenciar entre vídeos reales y vídeos *fake* (Priyaa et al., 2024).

---

6 PageRank es una medida sofisticada de centralidad que combina la idea de la importancia de los nodos con un modelo de navegación probabilístico, es decir, que simula el comportamiento de un usuario navegando aleatoriamente por la web y utiliza esa simulación para calcular la importancia de cada nodo, basándose en la cantidad y calidad de los enlaces que apuntan a él. La centralidad de eigenvector significa que la importancia de un nodo no sólo depende de cuántos otros nodos lo conectan, sino también de la relevancia de esos nodos.

7 Técnica de procesamiento de lenguaje natural que convierte el lenguaje humano en vectores matemáticos y es la base de los modelos de IA generativa, cuya respuesta con el número más alto es la más cercana a la pregunta.

8 Arquitectura RAG (Retrieval-Augmented Generation) es un modelo híbrido que combina los enfoques de recuperación de información y generación de texto. En este tipo de arquitectura, el sistema primero recupera información relevante de una base de datos o fuente externa y luego utiliza esa información como contexto para generar respuestas o contenido de manera más coherente y precisa. Esta integración permite que el modelo no sólo genere respuestas desde su conocimiento interno.

– Para el audio, se usará un modelo Whispers reconocido por su rendimiento de vanguardia en la conversión de audio a texto, logrando una alta precisión de transcripción (Haz et al., 2023) y posteriormente, una vez obtenido el texto, se procesará de manera análoga al texto del bulo.
– Como se ha puesto de manifiesto en el estudio de la fase dos, los bulos están redactados de manera no neutra (lenguaje directo y sensacionalista). Los algoritmos más efectivos en la actualidad para el análisis de estilo y sentimiento, debido a la capacidad de capturar el contexto bidireccional de las palabras en una oración, son BERT y RoBERTa. Para ello, se añadirán clasificadores. Por ejemplo, para el estudio propuesto se puede usar el clasificador sensacionalista, de forma que: 0 = no sensacionalista y 1 = sensacionalista, creando diferentes clasificadores para indicar el estilo directo y comprobar si producen o no sentimiento de polarización.

3.2.3. Verificación de hechos

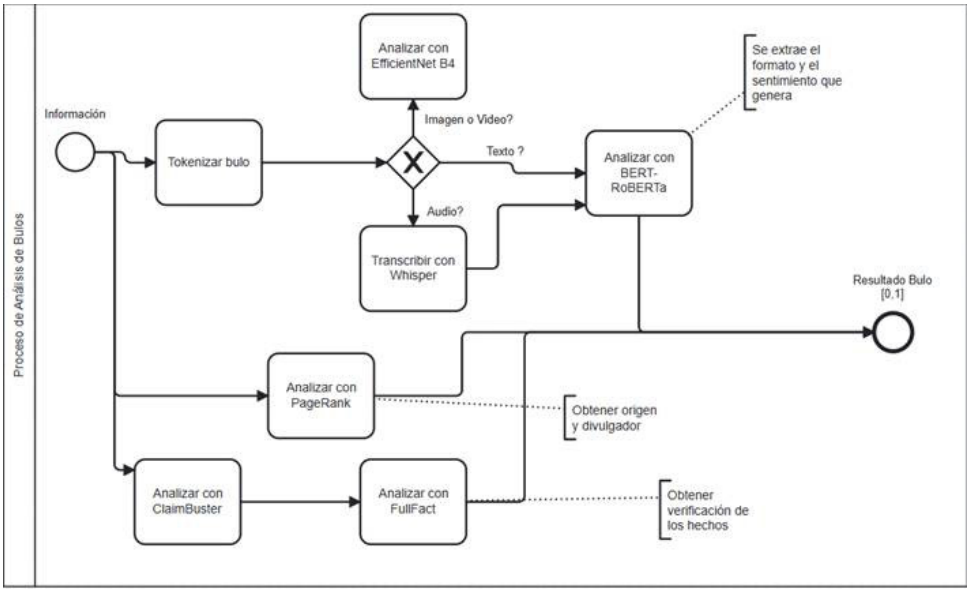
Debido a la rapidez con que se propagan los bulos y el corto plazo en que se desarrollan unas elecciones (en la ficha identificativa de bulos se puede observar que incluso existen el mismo día de las elecciones), será necesario que el sistema elegido pueda funcionar en tiempo real, que esté integrado con diferentes medios de comunicación y que contenga sistemas de alerta para verificadores humanos. Por tanto, se usarán ClaimBuster y Full Fact (ese último, además, aporta comprobación de discursos políticos y medios de comunicación específicos y cuenta con una base de datos de afirmaciones contrastadas). A su vez, para la extracción de hechos se recurrirá al algoritmo BERT-QA, siendo la *feature* más adecuada la contextualización, que se define como el proceso de comprobación contra dichas bases de datos.

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

3.3. Proceso

Imagen 1. Diagrama del proceso de análisis



Fuente: elaboración propia

El diseño del sistema propuesto se pude ver en la Imagen 1, que muestra el proceso que se describe a continuación, de manera que, una vez recibido un bulo, este se *tokeniza*[9], siendo la entrada para los siguientes subprocesos:

- Se preguntará por el formato de manera que:
  - Si es texto se analizará con BERT-RoBERTa.
  - Si es audio se transcribirá con Whisper y posteriormente se analizará con BERT-RoBERTa.
  - Si es vídeo se analizará con EfficientNet B4.
- Se extraerá la información de formato y sentimiento.
- Se analizará difusión con PageRank para obtener el origen y el divulgador.
- Se verificarán los hechos ClaimBuster y se confirmarán con Full Fact.

9 Un *token* es una unidad básica de texto que los mencionados algoritmos utilizan para procesar y analizar la información. Puede ser una palabra, una subpalabra, etc. La *tokenización* es el proceso de dividir el texto en estas unidades básicas.

Finalmente, se unificará toda la información de cada subproceso para dar una respuesta final binaria [0,1] para decidir si es un bulo o no.

Visto todo lo anterior, quedarían dos aspectos fundamentales para validar el sistema propuesto: el análisis de la sensibilidad y la especificad y la valoración de la explicabilidad.

3.4. Análisis de la sensibilidad y la especificidad

Las tasas de verdaderos positivos (TPR, True Positive Rate) y falsos positivos (FPR, False Positive Rate) son métricas claves utilizadas en la evaluación de modelos de clasificación, especialmente en problemas de clasificación binaria. Nuestro sistema pretende realizar la clasificación binaria ("bulo", "no bulo") de cada una de las informaciones de entrada (Jeni et al., 2013). Así:

– TPR, también conocida como sensibilidad. Mide la proporción de ejemplos positivos correctamente identificados por el modelo. Es una medida de cuán bien el modelo captura los casos positivos. El TPR se calcula como TP / (TP+FN), donde TP (True Positives) es el número de bulos correctamente clasificados como bulos y FN (False Positives) el número de bulos incorrectamente clasificados como "no bulos". Por lo tanto, cuanto más se acerque a 1 (todos los bulos son detectados) mejor será su sensibilidad.

– FPR, también conocida como especificidad. Mide la proporción de ejemplos negativos que el modelo clasifica incorrectamente como positivos. Es una medida de los errores cometidos con respecto a los ejemplos negativos. El FPR se calcula como FP / (FP+TN), donde FP (False Positives) es el número de "no bulos" clasificados erróneamente como "bulos" y TN (True Negative) es el número de "no bulos" clasificados correctamente como "no bulos". Por lo tanto, cuanto más se acerque a 0 (ninguna verdad es detectada como bulo) mayor especificad tendrá.

3.5. Explicabilidad

Aunque no forma parte del sistema en sí, la explicabilidad, es decir, la manera en la que el sistema explica cómo ha llegado a la conclusión de que una información es o no un bulo, es un componente importante para dotar al mismo de fiabilidad.

Para ello se usará SHAP (*SHapley Additive exPlanations*) que está basado en la teoría de juegos y proporciona explicaciones consistentes y aditivas. Esto significa que la suma de las contribuciones de todas las características da como resultado la predicción del modelo, lo cual es intuitivo y fácil de entender. SHAP se puede aplicar a cualquier tipo de modelo, tanto para interpretaciones locales (predicción individual) como globales (a lo largo de todo el conjunto de datos), algo crucial para comprender tanto casos específicos como tendencias generales en el contexto de noticias falsas de contenido político, ya que las explicaciones de SHAP pueden manejar interacciones complejas entre características. Esto último es relevante en el análisis de noticias falsas, donde múltiples factores (como el contenido textual, la temática, el autor, la fuente y la fecha) pueden interactuar de maneras no triviales.

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

4. Discusión y conclusiones

Tras la exposición de los resultados y el análisis previo, se puede decir que la IA posee un gran potencial en el ámbito periodístico, pero se rebela como un arma de doble filo. En los últimos años se ha manifestado su alta capacidad para generar contenidos falsos, utilizados de forma profusa con el fin de manipular a la opinión pública, especialmente en contextos electorales. Las consecuencias nocivas de la desinformación para el correcto funcionamiento de las democracias nos han conducido a una reflexión sobre la necesidad de emplear esta misma tecnología para combatir de forma eficaz la proliferación de bulos.

Partimos de una situación concreta, como es la precampaña y campaña de las últimas elecciones autonómicas en Cataluña, para analizar los patrones seguidos en la creación y difusión de noticias falsas. En línea con lo que corroboran trabajos como el de González-Quintero y Cardona-Restrepo (2023), ponemos el foco en las redes sociales como principal vía para la propagación de estos, en concreto en la red X.

El análisis de los catorce principales bulos detectados por Maldito Bulo (2024) que constituyen nuestra muestra y que consideramos representativo por ejemplificar un conjunto más amplio de este tipo de contenidos no registra ningún caso de *deepfake*, a pesar de ser esta, tal como afirma el informe de Home Security Heroes (2024), una tendencia creciente a nivel global. Observamos un predominio de supuestas noticias 100% falsas (diez de catorce), por lo que nuestros resultados no responden al esquema de tergiversación basado en informaciones parcialmente reales descrito por Alandete-Ballester (2019) (tan común, por otro lado, en este tipo de fenómenos).

Los temas más recurrentes son de índole social y en la mayoría de los casos conectados con la inmigración, aunque se presentan como noticias referidas a ámbitos diversos, entre ellos la educación, o las ayudas sociales. En este aspecto se manifiesta el sensacionalismo y el afán por desacreditar las instituciones a los que se refieren autores como Arrieta-Castillo & Rubio-Jordán (2023). Los ataques dirigidos principalmente a ERC y el partido socialista (a través de su división catalana), al frente respectivamente de los gobiernos de Cataluña y España, dan cuenta de este propósito, que se intensifica el mismo día de los comicios con el bulo sobre el sabotaje en la circulación de los Rodalíes. Por todo ello, se estima que contribuye a este fin el tono agresivo expuesto por nuestros datos, con predominio de signos de exclamación o letras mayúsculas en los textos publicados. Todo ello fomenta la crispación y la desconfianza e incrementa el riesgo para la estabilidad democrática, ilustrando uno de los riesgos señalados en el Libro Blanco contra la Desinformación (Gobierno de España, 2022).

La desinformación procede en once de los catorce ejemplos estudiados de perfiles que no pertenecen a alguien reconocible (como un político o cargo público) y que corresponderían, en apariencia, a cualquier ciudadano. En efecto, los formatos son sencillos (texto más fotografía en la mayoría de los casos) y sólo para las imágenes editadas se requiere un mínimo plus de destreza. Sin embargo, es probable que, bajo esos perfiles, en principio anodinos, se escondan activistas al servicio de uno u otro partido, siendo los afines a VOX los más activos. Fuera de esta pauta son los líderes de la ultraderecha (VOX y Partido Obrero) los que originan algunos de los bulos (dos) de nuestra muestra. El PP, por su parte, añade un tercero, el ya mencionado sobre el mal funcionamiento de los trenes locales durante la jornada electoral, que resulta el más difundido del conjunto observado.

A ello contribuyen los medios de comunicación, que no comparten ninguna de las noticias falsas recopiladas, pero, en este caso, recurren al *clickbait* y se hacen eco con titulares alarmistas de los cortes en los trenes de cercanías. Lejos de tratarse de una

conclusión positiva, ya que se detecta un único caso, este dato indica una práctica nefasta, que se suma al ruido y la confusión promovidos en la red X.

Evaluados estos resultados pasamos a elaborar una herramienta de IA que, basándose en las dinámicas detectadas en la fabricación y expansión de bulos, pueda identificar los intentos de desinformación relacionados con una convocatoria electoral.

Compartimos con Ruffo et al. (2023) la premisa de que no hay un sólo "mejor" algoritmo, por lo que combinamos los algoritmos y *features* más adecuados al foco de nuestro análisis. El sistema de IA propuesto está adaptado específicamente a la tipología de contenidos falsos de manera que, una vez que se han determinado los diferentes componentes de la misma (Tabla 3 y Tabla 4), se diseña un sistema que combina diferentes técnicas, algoritmos y *features*, adaptados a la muestra seleccionada. Analizamos los patrones de difusión con PageRank (en la red X) y, de acuerdo con las conclusiones de Zhang et al. (2004), ajustamos a nuestros intereses el tratamiento de la imagen con EfficientNet y de la información textual con un modelo de transformadores (BERT y RoBERTa), que además sirve para la realización del análisis de sentimiento (como por ejemplo el sensacionalismo). Una vez *tokenizada* la información, esta deberá ser comprobada con herramientas de *fact-checking*, como ClaimBuster o Full Fact, que tengan respuesta en tiempo real.

El sistema de IA propuesto pretende realizar la clasificación binaria ("bulo", "no bulo") de cada una de las entradas de información, a través del proceso descrito en la Imagen 1, por lo que será crucial medir la sensibilidad (proporción de ejemplos positivos correctamente identificados) y la especificidad (proporción de ejemplos negativos clasificados incorrectamente como positivos).

A su vez, y debido a que los algoritmos utilizados en este sistema lo permiten, se usará SHAP para realizar la explicabilidad de la decisión binaria del sistema, donde múltiples factores (como el contenido textual, la temática, el formato, el autor y la fecha) pueden interactuar de maneras no triviales.

Entre las limitaciones de esta investigación cabe señalar el tamaño de la muestra, ya que los catorce bulos seleccionados por Maldita, si bien presentan pautas que podrían extrapolarse a un conjunto mayor, no constituyen un conjunto numeroso. Como hemos señalado en la Metodología, la muestra constituye un conjunto de bulos escogidos de forma intencional en un contexto político concreto, no siendo su representatividad estadística o teórica.

Así mismo, es preciso en el diseño en desarrollo del recurso de IA refinar las características detectadas en los patrones de los bulos analizados y usadas como propiedades externas de cada algoritmo empleado en el proceso. Igualmente es necesario realizar un ajuste de parámetros en datos específicos sobre un conjunto más amplio de bulos relativos a otras elecciones autonómicas españolas, con el fin de entrenar el modelo y comprobar las métricas aplicadas para especificidad y sensibilidad en una muestra más amplia en el tiempo.

El recurso de IA planteado está adaptado a las características específicas de este tipo de contenidos (información falsa en procesos electorales en España), que pueden tener similitudes con otros ejemplos de desinformación, procedentes de situaciones y lugares diversos. Por ello, consideramos como objeto de las siguientes fases de estudio la aplicación de la precisión del modelo sobre nuevas citas electorales nacionales y su adecuación a dichos eventos celebrados fuera de nuestro país, a través de la medición de los parámetros de calidad propuestos y de la capacidad del sistema para realizar una autoexplicación de los resultados.

ISSN: 1696-019X / e-ISSN: 2386-3978

5. Agradecimientos

Este artículo ha sido traducido al inglés por Mario Fon a quien agradecemos su trabajo.

6.Fuentes de financiación

ISSN: 1696-019X / e-ISSN: 2386-3978

Contribuciones específicas de cada autor/a

	Nombre y apellidos
Concepción y diseño del trabajo	Mercedes Herrero de la Fuente, Celia Sancho Belinchón y Jorge Sedeño López
Metodología	Mercedes Herrero de la Fuente, Celia Sancho Belinchón y Jorge Sedeño López
Recogida y análisis de datos	Celia Sancho Belinchón
Discusión y conclusiones	Mercedes Herrero de la Fuente, Celia Sancho Belinchón y Jorge Sedeño López
Redacción, formato, revisión y aprobación de versiones	Mercedes Herrero de la Fuente

7. Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

8. Referencias bibliográficas

AIMC (2024). *Marco General de los Medios en España 2024*. https://cutt.ly/Kw2ghBxA

Alandete-Ballester, D. (2019) *Fake news: la nueva arma de destrucción masiva: Cómo se utilizan las noticias falsas y los hechos alternativos para desestabilizar la democracia*. Deusto.

Arrieta-Castillo, C., y Rubio-Jordán, A. V. (2023). Periodismo de verificación en formato vertical: narrativas multimedia de los verificadores en TikTok. *Ámbitos. Revista Internacional De Comunicación*, 60, 13-32. https://doi.org/10.12795/Ambitos.2023.i60.01

Badgamia, N. (1 mayo 2023). Explained. AI journalism: Can artificial intelligence replace journalists? *WioNews*. https://cutt.ly/Rw2gWt3x

Blanco-Marañón, N. (noviembre 2023). Qué diría Aristóteles de la inteligencia artificial. *Telos*, 123, 35-39. https://cutt.ly/6rWQCBxr

Boletín Oficial del Estado (BOE) (12 julio 2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial*. https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079

Cerezo-Guilarranz, P.(2024). Tendencias 2024. Hacia el final de la web abierta. *Programmatic Spain*. https://cutt.ly/Sw2gWQTs

Comisión Europea (2018). *A multi-dimensional approach to disinformation*. https://cutt.ly/Srjh4cNL

Criado-Grande, J.I. (2021). Inteligencia Artificial (y Administración Pública). *Eunomía. Revista en Cultura de la Legalidad*, 20, 348-372. https://doi.org/10.20318/eunomia.2021.6097

Deeptrace (2019). *The state of deepfakes. Landscape, threats and impact.* https://sensity.ai/reports/

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv,* 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Díez-Garrido, M., Renedo-Farpón, C., y Cano-Orón, L. (2021). La desinformación en las redes de mensajería instantánea. Estudio de las fake news en los canales relacionados con la ultraderecha española en Telegram. *Miguel Hernández Communication Journal*, *12*(2), 467-489. https://doi.org/10.21134/mhjournal.v12i.1292

Flores-Vivar, J. M. (2019). Inteligencia artificial y periodismo: diluyendo el impacto de la desinformación y las noticias falsas a través de los bots. *Doxa Comunicación*, 29, 197-212. https://doi.org/10.31921/doxacom.n29a10

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

Gobierno de España (2022). *Lucha contra las campañas de desinformación en el ámbito de la seguridad nacional. Propuestas de la sociedad civil*. https://cutt.ly/VrWQOXJc

González-Quintero, J. I., y Cardona-Restrepo, P.(2023). Post-truth and Social Networks as Challenges for Journalism in the Digital. *Ánfora, 30*(55), 332-359. https://doi.org/10.30854/anf.v30.n55.2023.977

Hashmi, E., Yayilgan, S.Y., Yamin, M.M., Ali, S., & Abomhara, M. (2024). Advancing Fake News Detection: Hybrid Deep Learning with FastText and Explainable AI. *IEEE Access*, 12, 44462-44480. https://doi.org/10.1109/ACCESS.2024.3381038

Haz, L., Fajrianti, E.D., Funabiki, N., & Sukaridhoto, S. (14-15 octubre 2023). *A Study of Audio-to-Text Conversion Software Using Whispers Model.* Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE), Surabaya (Indonesia). https://doi.org/10.1109/ICVEE59738.2023.10348186

Herrero-De-la-Fuente, M., y Ríos-Calvo, C. Construcción de un escenario para la posverdad: redes sociales y desinformación (2022). En A. Pérez-Escoda y J. Rubio-Romero (eds.). *Redes sociales, ¿el quinto poder? Una aproximación por ámbitos al fenómeno que ha transformado la comunicación pública y privada* (pp.79-97). Tirant lo Blanch.

Home Security Heroes (2024). *2023 State of Deepfakes.* https://cutt.ly/QrWQP1Cu

Jacobsen, B.N., & Simpson, J. (2023). The tensions of deepfakes. *Information, Communication Society, 27*(6), 1095-109. https://doi.org/10.1080/1369118X.2023.2234980

ISSN: 1696-019X / e-ISSN: 2386-3978

ISSN: 1696-019X / e-ISSN: 2386-3978

Jeni, L. A., Cohn, J.F., & De La Torre, F. (12 diciembre 2013). *Facing Imbalanced Data Recommendations for the Use of Performance Metrics.* Humaine Association Conference on Affective Computing and Intelligent Interaction, Ginebra (Suiza). https://doi.org/10.1109/ACII.2013.47

López-López, P.C., Lagares-Díez, N., y Puentes-Rivera, I. (2021). *Razón y Palabra*, *24*(111), 5-11. https://razonypalabra.net/index.php/ryp/article/view/1891/1681

Lopezosa-García, C., Pérez-Montoro, M., y Rey-Martín, C. (2024). El uso de la inteligencia artificial en las redacciones: propuestas y limitaciones. *Revista de Comunicación*, *23*(1), 279-293. https://doi.org/10.26441/RC23.1-2024-3309

Luengo-Cruz, M., y García-Marín, D. (2020). The performance of truth: politicians, fact-checking journalism, and the struggle to tackle COVID-19 misinformation. *American Journal of Cultural Sociology*, 8, 405-427. https://doi.org/10.1057/s41290-020-00115-w

Maldito Bulo (17 mayo 2024). *14 bulos y desinformaciones sobre las elecciones autonómicas Cataluña del 12 de mayo de 2024.* https://cutt.ly/uenhODjL

Moreno-Espinosa, P., Abdulsalam-Alsarayreh, R. A., y Figuereo-Benítez, J. C. (2024). El Big Data y la inteligencia artificial como soluciones a la desinformación. *Doxa Comunicación*, 38, 437-451. https://doi.org/10.31921/doxacom.n38a2029

Newman, N., Fletcher, R., Eddy, K., Robertson, C.T., & Nielsen, R.K. (2024). *Reuters Institute Digital News Report 2023.* Reuters Institute, University of Oxford. https://cutt.ly/6rWQSU6x

Peña-Fernández, S., Meso-Ayerdi, K., Larrondo-Ureta, A, y Díaz-Noci, J. (2023). Sin periodistas, no hay periodismo. La dimensión social de la inteligencia artificial generativa en los medios de comunicación. *Profesional de la información*, *32*(2), e320227. https://doi.org/10.3145/epi.2023.mar.27

Priyaa, V. G., Harrish, M. J., Udhayakumar, M., Jothieswaran, N., & Suresh, S. (12-14 abril 2024). *EfficientNet-Based Deep Learning Approach for Video Forgery Detection and Authentication.* 10th International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur (India). https://doi.org/10.1109/ICCSP60870.2024.10544113

Reinsel, J., Gantz, J., & Rydning, D., (2018). *The digitization of the world from edge to core*. International Data Corporation. https://cutt.ly/frWQDLdu

Rodríguez-Martelo, T., Rúas-Araújo, J., y Maroto-González, I. (2023). Innovation, digitization, and disinformation management in European regional television stations in the Circom network. *Profesional de la información, 32*(1). https://doi.org/10.3145/epi.2023.ene.12

Ruffo, G., Semeraro, A., Giachanou, A., & Rosso P. (2023) Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, 47. https://doi.org/10.1016/j.cosrev.2022.100531

Sádaba, C., y Salaverría, R. (2023). Combatir la desinformación con alfabetización mediática: análisis de las tendencias en la Unión Europea. *Revista Latina de Comunicación Social*, 81, 17-33. https://www.doi.org/10.4185/RLCS-2023-1552

Simanjuntak, A., Lumbantoruan, R., Sianipar, K., Gultom, R., Simaremare, M., Situmeang, S., & Panggabean, E. (2024). Research and Analysis of IndoBERT Hyperparameter Tuning in Fake News Detection. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *13*(1), 60-67. https://doi.org/10.22146/jnteti.v13i1.8532

Telefónica Tech (26 enero 2018). *Machine Learning contra "fake news".* https://cutt.ly/RrWQLvLr

UNESCO (2021). *TVETipedia Glossary.* UNESCO International Centre for Technical and actional Education and Training. https://cutt.ly/Ew2grz3A

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., & Yu, P.S. (2018). TI-CNN: Convolutional neural networks for fake news detection. *arXiv*,1806.00749. https://doi.org/10.48550/arXiv.1806.00749

Zhang, Z., Lv, Q., Jia, X., Yun, W., Miao, G., Mao, Z., & Wu, G. (2024). GBCA: Graph Convolution Network and BERT combined with Co-Attention for fake news detection. *Pattern Recognition Letters*,180, 26-32. https://doi.org/10.1016/j.patrec.2024.02.014

ISSN: 1696-019X / e-ISSN: 2386-3978