VISUAL REVIEW | Vol. 16, No. 4, 2024 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v16.5303

IMPLEMENTACIÓN DE HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL EN LA DETECCIÓN DE VÍDEOS FALSOS Y ULTRAFALSOS (DEEPFAKES)
Caso de Radio Televisión Española (RTVE)

MARTA SÁNCHEZ ESPARZA [1], SANTA PALELLA-STRACUZZI [2], ÁNGEL FERNÁNDEZ FERNÁNDEZ [3]
[1] Universidad Internacional de la Empresa (UNIE), España
[2] EAE Business School_Madrid, España
[3] Centro Universitario The Core Entertainment Science School, España

PALABRAS CLAVE

Falso
Deepfakes Vídeos
Inteligencia artificial
Radio Televisión Española

RESUMEN

La preocupación por la difusión de información falsa ha llevado a medios a emplear la inteligencia artificial (IA) para detectar deepfakes. Esta investigación es descriptiva-exploratoria. Mediante una revisión bibliográfica y entrevistas, revela el impacto transformador de la IA destacando su empleo para verificar la autenticidad de los contenidos. En este ámbito RTVE combina metodologías tradicionales con otras basadas en IA, y lidera el desarrollo de varias herramientas en colaboración con diferentes universidades. Estas herramientas han dado ya resultados satisfactorios en la detección de estos materiales, fortaleciendo la veracidad de la información y aumentando la confianza ciudadana en sus contenidos.

Recibido: 17 / 04 / 2024
Aceptado: 05 / 06 / 2024

1. Introducción

La desinformación, entendida como la difusión intencionada y, por lo general encubierta, de información falsa con el objetivo de manipular a la opinión pública o desestabilizar a la audiencia (Fernández et al., 2020) ha experimentado una transformación muy significativa en la era digital. Aunque la mentira y la propaganda han sido utilizadas de manera constante a lo largo de la Historia como herramientas de control político y social (Pineda, 2004), resulta evidente que nunca antes la humanidad había tenido a su alcance herramientas tan potentes y eficaces como las actuales para diseminar bulos, mentiras y otros contenidos falsos. Cada período histórico ha empleado su propia tecnología para difundir información falsa y propaganda. Sin embargo, en la actualidad, el desarrollo de las tecnologías de la información ha transformado por completo este panorama, poniendo en circulación una cantidad de desinformación sin precedentes. La popularización de Internet y el aumento de la participación de los usuarios en la creación, clasificación y distribución de toda clase de contenidos digitales han propiciado que nuestra actividad diaria se haya convertido en un incesante intercambio de información y datos (Fernández, 2017). Si bien la mayor parte de estos contenidos no plantean ningún problema de veracidad, cada vez es más común que los usuarios opten por compartir contenidos total o parcialmente falsos que refuerzan sus propios sesgos y opiniones (Olmo y Romero, 2019).

Sin duda, una de las consecuencias más preocupantes de esta situación, es la devaluación del propio concepto de verdad. Aunque, en términos generales, es posible interpretar este fenómeno como una consecuencia del relativismo que caracteriza al pensamiento posmoderno, Kavanagh y Rich (2018) consideran que tras este fenómeno residen tendencias sociales cada vez más arraigadas en la actual cultura digital, como el creciente desacuerdo entre la interpretación subjetiva de los hechos y el análisis objetivo de los datos que los sustentan; la influencia cada vez mayor que ejercen la opinión y la experiencia personal en nuestra interpretación de la realidad o la desconfianza cada vez más profunda hacia los medios de comunicación y las fuentes de información objetiva que antes se consideraban legítimas.

En este contexto, el papel que asume la imagen resulta particularmente crítico. Debido a su capacidad de proyección simbólica, las imágenes plantean un modelo de comunicación eminentemente emocional, lo que las convierte en un medio privilegiado para la difusión de contenidos falsos y desinformación (Hameleers et al., 2020). Las imágenes resultan especialmente útiles para captar la atención de unas audiencias abrumadas por la sobreinformación, aportando verosimilitud y credibilidad a las informaciones falsas. Especialmente en las redes sociales, se evidencia que los contenidos que incluyen imágenes tienden a propagarse más que los que tienen un carácter exclusivamente textual. Así, por ejemplo, los tuits que contienen fotografías o vídeos obtienen alrededor de un 18% más de clics, un 89% más de «me gusta» y un 150% más de *retuits* que aquellos formados únicamente por texto (Cao et al.,2020).

La capacidad que proporciona la inteligencia artificial para crear vídeos ultrafalsos o *deepfakes* afecta profundamente a la estructura misma de los medios digitales. Aunque las capacidades finales de la inteligencia artificial aún están por concretarse y todavía no existe apenas regulación al respecto, es evidente que la intersección de los medios digitales y la inteligencia artificial es un elemento central en la evolución de la comunicación mediática actual. A través de los *deepfakes* y otros tipos de contenido generados por sistemas de inteligencia artificial, se desdibujan las fronteras entre lo real y lo falso, socavando la confianza del público en la información que recibe. Con los *deepfakes*, la capacidad de distorsionar la realidad ha experimentado una transformación radical. Por esta razón, resulta crucial investigar esta intersección desde diversas perspectivas.

En este contexto es relevante analizar cómo un medio de comunicación público como Radiotelevisión española afronta la proliferación de estos contenidos desinformativos, y como emplea las nuevas herramientas basadas en inteligencia artificial para detectar estos vídeos falsos y *deepfakes*.

El ente Radiotelevisión Española es una empresa pública con 6.600 empleados que está viviendo en estos momentos un importante proceso de transición digital. En dicho proceso la inteligencia artificial está siendo el principal catalizador del cambio. Entre los principales usos de esta tecnología se encuentra precisamente el análisis de contenido y la detección de imágenes y audios falsificados mediante IA.

2. Objetivos y Metodología

El objetivo general de este estudio es describir la implementación de herramientas de Inteligencia Artificial (IA) en Radio Televisión Española (RTVE), y cómo éstas se emplean en la detección de vídeos

falsos y ultrafalsos o *deepfakes*. Entendemos implementación en el sentido de incorporación de estas herramientas a los procesos de trabajo de la compañía.

A tal fin se busca dar respuesta a los siguientes objetivos específicos:

- Describir cómo se ha implementado la IA en RTVE.
- Conocer las herramientas que se utilizan de IA en diferentes áreas de RTVE.
- Explorar el uso de las herramientas de IA en la detección de vídeos ultrafalsos o *deepfakes* en RTVE.

Desde el punto de vista metodológico, se ha llevado a cabo una revisión bibliográfica y una investigación de campo de carácter exploratorio-descriptivo, estudiando el caso de Radio Televisión Española (RTVE). Para el trabajo de campo se ha utilizado la técnica de la entrevista, que permitió recolectar la información para dar respuesta a los objetivos.

El instrumento utilizado fue un guion de entrevista estructurada. El guion se construyó con 11 preguntas diferentes de acuerdo con el rol desempeñado por el entrevistado dentro de RTVE. La validez del guion de entrevista se verificó a través del juicio de tres expertos quienes verificaron que con las preguntas se daba respuesta a los objetivos planteados (Palella y Martins, 2017:106).

En tal sentido, entrevistamos al director de Innovación y Digital de RTVE, Urbano García, al director de Estrategia Tecnológica de RTVE, Pere Vila, y al director del Servicio Verifica de RTVE, Borja Díaz-Merry. Estas entrevistas se efectuaron entre el 17 y el 22 de noviembre del 2023.

La muestra específica de profesionales de RTVE se seleccionó bajo el criterio del nivel de aportación de información relevante sobre la implementación de herramientas de Inteligencia Artificial (IA) en un medio de comunicación, y de forma más concreta sobre el uso de esas nuevas herramientas en la detección de vídeos o ultrafalsos *deepfakes* y vídeos falsos en el caso de Radio Televisión Española (RTVE).

3. Revisión de la Literatura

3.1. La inteligencia artificial en los medios de comunicación

La inteligencia artificial (IA) es una rama de la informática que se centra en la creación de máquinas inteligentes capaces de realizar tareas que antes realizaban los humanos. En la industria de los medios, se está implementando cada vez más, lo que genera cambios y desafíos importantes.

La IA se utiliza en los medios para optimizar y mejorar operaciones, como el análisis de datos y la generación de contenido multimedia (Sančanin y Penjišević, 2022). Se puede utilizar también para automatizar procesos, incluyendo la gestión de las redes sociales, donde se pueden entrenar algoritmos para analizar las acciones, preferencias y reacciones de los usuarios (Al Husseiny, 2023).

La IA también se utiliza en la industria de las noticias para alterar los enfoques tradicionales, aprovechando el aprendizaje automático, planificando, programando y optimizando procesos, cada vez más desarrollados (Kalinová, 2022). La implementación de la IA en las plataformas de redes sociales se está volviendo inevitable, con aplicaciones que incluyen *chatbots*, que detectan comportamientos dañinos, analizan datos y elaboran estrategias (De Lima-Santos y Wilson, 2022). Puede decirse que la IA tiene el potencial de transformar las empresas de este sector y sus funciones (Sadiku et al., 2021).

Esta transformación supone para los profesionales de los medios una liberación de las tareas rutinarias que les permite producir contenidos de mayor calidad. Sin embargo, también genera preocupación por la creciente dependencia de las plataformas tecnológicas y la amenaza para la independencia editorial. Los trabajadores de los medios perciben una amenaza para sus puestos de trabajo y una pérdida potencial de su capital simbólico como intermediarios entre la realidad y las audiencias (Peña-Fernández et al., 2023).

Por todo lo anterior, la implementación de la IA en los medios plantea desafíos sociales y epistemológicos para los periodistas y la profesión. Igualmente existe un debate sobre el uso de estas tecnologías en los medios en el seno de la Unión Europea. En este ámbito los marcos regulatorios relacionados con la IA rara vez incluyen a los medios; cuando lo hacen, abordan cuestiones como la desinformación, los datos, la alfabetización en IA, la diversidad, la pluralidad y la responsabilidad social (Porlezza, 2023).

La Unión Europea, además, está librando una batalla contra los contenidos desinformativos, mediante el lanzamiento de diferentes planes estratégicos y la creación de grupos de trabajo. A juicio de la Comisión Europea, la desinformación no es sólo un problema colateral, sino un ecosistema (Jerónimo y Sánchez-Esparza, 2022).

Para hacer frente a los contenidos que se generan en dicho ecosistema, los periodistas dependen de una combinación de métodos tradicionales y digitales. Un estudio elaborado por Haidar (2023) destaca la dependencia de los informadores de sitios web y de herramientas gratuitos (69,2%). Si la importancia de la IA en el periodismo y los medios es incuestionable en el análisis y la creación de contenidos, su papel en la verificación de la información es clave.

3.2. Los vídeos falsos y ultrafalsos (deepfakes). Semejanzas y diferencias

Los vídeos falsos son aquellos que han sido manipulados o generados utilizando tecnología de inteligencia artificial para crear contenido que no es real o exacto. Se han desarrollado varias técnicas para detectar estos vídeos falsos, como métodos basados en el aprendizaje multimodal que combinan información de audio, vídeo y fisiología (Stefanov et al., 2022). También se han utilizado técnicas forenses basadas en biometría. Estas técnicas tienen como objetivo identificar discrepancias o anomalías en los vídeos que indiquen manipulación o falsificación (Matthews, 2023; Timothy y Shih, 2011).

Entre los contenidos desinformativos resultan especialmente peligrosos los vídeos manipulados, y en mayor medida los *deepfakes*, que permiten superponer el rostro de una persona sobre el cuerpo de otra, creando así un contenido falso y convincente. La detección de estos *deepfakes* es un desafío importante y entraña cada vez más dificultades, debido al rápido avance en las técnicas de manipulación facial (Al-Khazraji et al., 2023).

En su definición más común, los *deepfakes* abarcan fotografías, vídeos y audios generados digitalmente mediante técnicas de inteligencia artificial (Bañuelos, 2022) que representan de manera realista a individuos realizando acciones o expresando palabras que nunca han llevado a cabo o dicho (Cerdán y Padilla, 2019). Se trata, por tanto, de contenidos concebidos expresamente para generar información falsa y engañosa. Aunque, como acabamos de indicar, el término puede aplicarse a diversos formatos, su uso se ha vuelto cada vez más específico, refiriéndose principalmente a vídeos creados digitalmente. En estos vídeos, la cara y/o la voz de una persona se sobreponen a contenido previamente grabado por otra persona, o se fusionan con una imagen generada digitalmente utilizando técnicas de Machine Learning y Deep Learning.

Es importante destacar este matiz, ya que, a pesar de que la mayoría de los vídeos ultrafalsos o *deepfakes* consisten en la superposición del rostro de una persona sobre el cuerpo de otra, esta categoría también engloba imágenes y sonidos completamente nuevos, generados directamente mediante la síntesis de grandes conjuntos de datos por parte de sistemas de IA, sin partir necesariamente de una imagen o sonido real previo (Karnouskos, 2020).

Aunque los *deepfakes* se nutren de sistemas de inteligencia artificial basados en tecnología compleja, con frecuencia pueden ser creados mediante herramientas de fácil acceso y servicios disponibles para el público en general. De hecho, la mayoría de las herramientas utilizadas actualmente para generar *deepfakes* requieren pocos requisitos técnicos, pudiendo ser empleadas sin inconvenientes en ordenadores domésticos convencionales equipados con tarjetas gráficas de nivel medio.

La accesibilidad de estas tecnologías y la baja curva de aprendizaje de las herramientas más comúnmente empleadas para crear *deepfakes* constituyen uno de los mayores riesgos derivados de este tipo de contenidos. El hecho de que usuarios con escasos conocimientos técnicos puedan crear imágenes falsas extremadamente realistas, favorece la generación y difusión de este tipo de contenidos, aumentando lógicamente los riesgos derivados de su utilización. Cuando consideramos la facilidad con la que los *deepfakes* pueden ser distribuidos a través de las redes sociales, resultan evidentes las graves implicaciones que pueden tener en la propagación de bulos y otras formas de desinformación.

Han ido apareciendo, sin embargo, herramientas igualmente generadas con inteligencia artificial que persiguen la detección de estos vídeos falsos. Así, a las metodologías desarrolladas por algunos expertos desde la perspectiva de los periodistas (Sohrawardi et al., 2019), se han añadido trabajos sobre la efectividad de las redes neuronales (Shilpa et al., 2023). Estas herramientas pueden

distinguir entre rostros reales y falseados gracias a modelos de entrenamiento (Haseena et al., 2023).

Los vídeos falsos y los *deepfakes* comparten similitudes, pues ambos implican la creación de contenido manipulado y pueden usarse para alterar la apariencia o las acciones de las personas, lo que genera posible desinformación y distorsión de la verdad (Matthews, 2023).

Los *deepfakes* plantean desafíos únicos, ya que pueden generar un alto grado de riesgo epistémico, lo que podría generar escepticismo sobre el conocimiento de los vídeos en línea (Liz-López, 2023) En general, los *deepfakes* representan una forma más avanzada y sofisticada de manipulación, y exigen una inversión mucho mayor en técnicas y tiempo de elaboración (Díaz-Merry, B., conversación en entrevista el 22-11-2023).

4. Resultados

4.1. Implementación de Inteligencia Artificial en RTVE

La Corporación Pública Radiotelevisión Española (RTVE) lleva años integrando tecnologías de Inteligencia Artificial (IA) en diferentes procesos y departamentos. Ya en el año 2015 se inició un programa de investigación en torno a las oportunidades que ofrecen los sistemas de procesamiento inteligente de la información (Aramburú et al., 2023). Este programa reúne a expertos en procesos de IA, estudiantes y profesores, y se sustenta mediante iniciativas como la Cátedra RTVE-UAB (con la Universidad Autónoma de Barcelona) o el Observatorio para la Innovación de los Informativos en la Sociedad Digital (OI2).

Puede además considerarse una fecha clave la llegada en 2021 a la dirección de RTVE del catedrático de Periodismo de la Universidad Autónoma de Barcelona José Manuel Pérez Tornero, quien decidió crear la Dirección de Innovación y Digital, poniendo al frente al periodista Urbano García. Se trata de un área dirigida a acometer un plan estratégico para transformar una televisión donde se elaboraban algunos contenidos digitales en una empresa «de core completamente digital entre cuyas actividades está la televisión» (García, U., comunicación personal, 17 de noviembre de 2023).

Esta nueva Dirección de Innovación y Digital ha absorbido el Observatorio para la Innovación OI2, así como la gestión de las cátedras y el Laboratorio de Innovación Audiovisual (Lab) de RTVE, un departamento dedicado a explorar nuevas narrativas. Además, esta dirección es responsable de la estrategia relacionada con los nuevos medios y de la transición de toda la empresa al nuevo modelo digital, alineando los valores que conforman la misión de la televisión pública con la utilización de nuevas tecnologías como la IA para la elaboración de contenidos.

En este proceso de transición las tecnologías basadas en IA están impactando ya en la forma de hacer televisión y en el propio modelo de negocio. Ante los desafíos que comporta, los responsables de RTVE han acometido una reflexión general sobre los límites, riesgos y oportunidades que supone, junto a otras empresas públicas como la Sociedad Estatal de Participaciones Industriales (SEPI).

A juicio de Pere Vila, director de Estrategia Tecnológica de RTVE, la inteligencia artificial va a infiltrarse en todas las actividades de la compañía, no solo en áreas como la documentación -mediante el metadatado de todos los archivos de los fondos de RTVE-, sino en proyectos de análisis y automatización de contenido, tratamiento y coloreado de imágenes, clonación de voces, generación de avatares, selección de personas o relación con las audiencias y recomendación de contenidos, entre otros (Vila, P., comunicación personal, 22 de noviembre de 2023).

En la actualidad se trabaja en RTVE con tecnologías de IA en los ámbitos que se muestran en la tabla siguiente:

Tabla 1. Ámbitos de las tecnologías de Inteligencia Artificial en RTV

Análisis de contenido	Generación de contenido	Otras aplicaciones
Se utilizan tecnologías de reconocimiento de voz para generar subtítulos automáticos en tiempo real durante la emisión de programas en vivo. Se utiliza inteligencia artificial para la indexación automática del archivo documental, lo que permite organizar y etiquetar la información de manera más eficiente y precisa. Otro uso es el análisis y recuento de temas tratados o el contenido de vídeo, algo que sirve para elaborar informes de Responsabilidad Social Corporativa o, por ejemplo, para cuantificar el tiempo en el que se utiliza la lengua de signos. Se emplean sistemas de recomendación basados en los intereses de los usuarios, ofreciendo contenidos afines a sus preferencias.	Se utilizan datos e información previamente procesada para generar textos, gráficos y audios de manera automática. Creación de voces con inteligencia artificial que puedan hablar de forma natural, como si fuera una voz humana.	Aumento en la calidad de las imágenes de archivo, eliminando ruido, coloreándolas y mejorando su nitidez. Tecnología aplicada a proyectos contra la desinformación y nuevas formas de verificación. Creación de avatares personalizados, creados íntegramente con inteligencia artificial.

Fuente: Elaboración propia.

4.2. Herramientas de IA en diferentes áreas de RTVE

Por lo general, las herramientas empleadas son externas y se adquieren mediante la compra de licencias de uso. Una vez adquiridas, el equipo de tecnología de la corporación efectúa test para comprobar qué tipo de usos puede darse a esas aplicaciones en el trabajo diario de RTVE.

A continuación, en la siguiente tabla, incluimos las principales herramientas basadas en tecnologías de IA que se están utilizando en RTVE:

Tabla 2. Herramientas de IA en la RTVE

Herramientas de IA	Uso
Lexica	Creación de imágenes y detección de imágenes generadas con IA
Stable Difussion	Creación de imágenes de alta calidad a partir de textos
ChatGPT	Generación de textos y otros contenidos.
Dall-E	Creación de imágenes
HeyGen	Generación de avatares
Studio D-ID	Generación de avatares
Eleven Labs	Clonación de voces
Runway y Stable Diffusion	Transformación de imágenes y videoclips
Adobe (funcionalidades)	Relleno generativo
Caja de herramientas de acceso abierto	Verificación de información y deepfakes
Herramientas proyecto IVERES	Transcripción y traducción, detección de audios y vídeos falsos

Fuente: Elaboración propia.

4.2.1 La tecnología detrás de las herramientas de IA aplicadas en las áreas de RTVE

Detrás de las herramientas usadas para detectar vídeos falsos y *deepfakes* figuran las siguientes tecnologías:

- Reconocimiento facial y de voz: El análisis de las características de un rostro mediante el uso de algoritmos de reconocimiento facial compara la imagen analizada con imágenes de rostros conocidos almacenados en una base de datos, buscando inconsistencias (Guarnera y Battiato, 2023). En el caso del reconocimiento de voz, se analizan el timbre, tono y pronunciación, y se comparan con las voces almacenadas en las bases de datos.
- Análisis de metadatos: Los programas para modificar imágenes, vídeos y fragmentos de sonido introducen sus propios metadatos en el archivo o modifican los metadatos existentes

(fecha y hora de creación o modificación, software utilizado, ubicación, etc.). Nuevamente, la inconsistencia de la información puede revelar un d*eepfake.*

- Análisis forense: Se buscan inconsistencias en la luz, sombras, reflejos y pequeñas perturbaciones a nivel de píxeles. En el caso de las imágenes, se utilizan técnicas como el Análisis de Nivel de Error (ELA) (Martín-Rodríguez et al., 2023) o sistemas como *ProtoExplorer* (Bouter et al., 2023). Si es un vídeo, el análisis también abarca movimiento y perspectiva o expresiones faciales. En el caso del audio, los objetos de análisis suelen ser la forma de onda y el espectrograma (rango de frecuencia, armónicos, ruido de fondo, etc.).
- Aprendizaje automático e inteligencia artificial: Así como una red neuronal puede entrenarse para generar materiales multimedia, también pueden entrenarse para detectar la generación sintética de estos mismos materiales. Es decir, una red neuronal puede utilizarse para detectar la intervención de otra red neuronal. Al igual que cualquier otra red neuronal profunda, el entrenamiento de estos sistemas requiere conjuntos de datos grandes que incluyan imágenes reales y manipuladas.
- Tecnología Blockchain: Al igual que en cualquier otro ámbito donde se deba garantizar la imposibilidad de una modificación posterior de una transacción, esta tecnología puede usarse para la generación y garantía de autenticidad de un elemento multimedia. Almacenar esos elementos en un sistema blockchain hace virtualmente imposible alterarlos o manipularlos sin ser detectados. Blockchain puede garantizar el origen y la trazabilidad de los datos, creando una plataforma segura para almacenar e intercambiar información multimedia (Rashmi et al., 2023).

4.3. Casos de uso de Inteligencia Artificial en la creación de contenidos

Entre los productos que se han alumbrado en RTVE gracias al uso de IA figura el proyecto RTVEIA (www.rtveia.es), una web desde la que en la tarde de las últimas elecciones generales de 2023 se lanzaron 70.000 piezas informativas con texto, imágenes y voces sintéticas, informando en tiempo real de los resultados en casi 5.000 municipios españoles menores de 1.000 habitantes. A juicio de Urbano García, este tipo de contenidos refuerzan el papel de RTVE como servicio público mediante acciones de vertebración territorial en lugares donde no existen medios de comunicación. Allí la IA está permitiendo prestar un mejor servicio informativo, sin sustituir a los periodistas. El proyecto, que cuenta con la participación de la empresa Narrativa, Monoceros Labs, Universidad de Castilla-La Mancha, Universidad de Granada, ONCE, AWS y el Centro Territorial de Castilla-La Mancha de RTVE, recibió el premio IBC 2023 a la Innovación y el Impacto Social en septiembre de 2023.

Otro de los productos desarrollados desde RTVE mediante IA es RTVE2030 (www.rtve2030.rtve.es), una web donde se analizan cada día los contenidos de los programas de actualidad y se mide el tiempo dedicado a cada uno de los Objetivos de Desarrollo Sostenible (ODS). Con esas analíticas (figura 1) se elaboran más tarde los informes sobre políticas de RSC de la corporación.

Figura 1. Web de RTVE2030, con analíticas sobre los ODS mediante IA

Fuente: www.rtve2030.rtve.es

Un tercer proyecto, lanzado en febrero de 2023, es Hiperia, un avatar generado con inteligencia artificial que presenta un espacio semanal sobre música y cultura juvenil para Radio 3. El personaje ha sido diseñado gracias a la colaboración entre Radio 3 y las áreas de Estrategia Tecnológica, Innovación y Digital y Grafismo. Tanto el personaje como su voz, el guion y los contenidos del programa están hechos mediante inteligencia artificial (Vila, P., conversación personal, 22-11-2023).

Figura 2. Imagen de Hiperia, un presentador creado con IA para un espacio en Radio 3. Fuente:

www.rtve.es/radio

4.4. El uso de IA en la detección de vídeos falsos y ultrafalsos o deepfakes en RTVE

4.4.1 El servicio Verifica RTVE

En el año 2020, la dirección de RTVE creó de manera formal el servicio Verifica RTVE, integrado por cuatro periodistas. La corporación había estado trabajando anteriormente para combatir la desinformación, pero fue en el transcurso de la pandemia cuando se resolvió crear un servicio específico ante el exceso de bulos e información errónea (Díaz-Merry, B., conversación personal, 22-11-2023).

El equipo de Verifica RTVE cuenta en la actualidad con seis periodistas, todos titulados en Ciencias de la Información y capacitados, mediante una formación específica impartida dentro de RTVE, para llevar a cabo las tareas de verificación de información e investigación. Estos profesionales se dedican a la verificación del discurso político ligado a grandes eventos, como los debates parlamentarios o los debates en el transcurso de una campaña electoral, en coordinación con los periodistas de los servicios informativos.

Los periodistas de Verifica RTVE trabajan de manera proactiva en la monitorización de redes sociales, en busca de mensajes sospechosos, fraudulentos, falsos o engañosos. Y en un segundo nivel, trabajan a nivel interno, a demanda de los servicios informativos de RTVE, que necesitan cotejar la veracidad de vídeos, fotos, informes y documentos antes de utilizarlos en las informaciones que serán emitidas ese día.

Según el responsable de Verifica RTVE, Borja Díaz-Merry, es esa línea de trabajo interna la que crece a mayor velocidad y exige del equipo mayores esfuerzos, especialmente desde el estallido de la guerra de Ucrania, en febrero de 2022. Si antes del conflicto la demanda interna de peticiones de verificación de materiales desde los servicios informativos era de 2 o 3 al mes, a partir de la invasión rusa el ritmo de peticiones se disparó hasta las 2 o 3 peticiones diarias, tanto para las dos ediciones de Telediario como para el canal 24 horas de RTVE.

Desde ese momento la dimensión de verificación interna tomó un papel preponderante en el trabajo de Verifica RTVE, chequeando vídeos a demanda de los telediarios, y también para los servicios territoriales de RTVE, donde llegan vídeos de inundaciones, operaciones policiales, sucesos, etcétera, de muy diversa índole. El equipo sigue analizando los contenidos falsos que más se viralizan en redes sociales para verificarlos, pero el grueso de su labor ha pasado a ser el trabajo al servicio de la fiabilidad de los informativos de la emisora pública.

En ese marco, el servicio suele detectar entre tres y cuatro historias diarias viralizadas en redes sociales, de las que se analizan algunas en función de criterios periodísticos, y dedicando más tiempo a estas historias, al no tener que ser publicadas en el día. En cambio, las peticiones de los servicios

informativos deben ser resueltas a mayor velocidad. Según Díaz-Merry, el volumen de verificaciones que llevan a cabo a petición interna de los servicios informativos es de unas veinte mensuales.

En este capítulo es preciso distinguir entre vídeos falsos -conocidos como *shallow fakes*- y los llamados *deepfakes*. Mientras que los primeros son vídeos manipulados mediante una edición más sencilla, los *deepfakes* son más sofisticados y elaborados. Se trata de creaciones digitales en la que hay una grabación real sobre la que se superponen millones de fotografías para suplantar el rostro de una persona. En algunos casos, lo que se suplanta es la voz, mediante el uso de la misma tecnología.

Según Díaz-Merry, esta sofisticación en la elaboración hace que los *deepfakes* no sean tan frecuentes como los vídeos simplemente falsos y manipulados. De hecho, mientras que desde Verifica RTVE suelen detectarse dos o tres *shallow fakes* al mes, los *deepfakes* no aparecen con tanta frecuencia. No todos estos vídeos se verifican, pues desde RTVE se practica la 'verificación responsable', es decir, se estudia el peligro que entraña ese contenido y su grado de viralización, y se resuelve si es mejor publicar su verificación o simplemente advertir a los servicios informativos, para evitar darle mayor difusión al tema.

4.4.2. Herramientas de IA en la detección de vídeos falsos y ultrafalsos o deepfakes en RTVE

Para analizar estos vídeos falsos, los profesionales utilizan herramientas de IA unidas al análisis periodístico. La mayor parte de estas herramientas son de acceso público y gratuitas, y se ofrecen a los usuarios en las llamadas 'cajas de herramientas' de Verifica RTVE, incluidas en su página web. Se trata de dos cajas de herramientas, una básica y otra avanzada, disponibles para que cualquier persona pueda chequear informaciones falsas.

Por otra parte, RTVE lidera junto a la Universidad Autónoma de Barcelona el proyecto IVERES (Investigación, Verificación y Respuesta), dotado con fondos europeos Next Generation, en el que también participan la Universidad de Granada, la Universidad Politécnica de Barcelona y la Universidad Carlos III, y donde se están desarrollando cajas específicas de herramientas para Verifica RTVE que utilizan inteligencia artificial. Se trata de tres tipos de herramientas: herramientas de transcripción y archivo, herramientas de detección de audios falsos y herramientas de detección de vídeos falsos y ultrafalsos. Cada una de ellas es desarrollada desde una de las tres universidades participantes.

Dentro del proyecto IVERES, la herramienta desarrollada y monitorizada desde la Universidad Carlos III para el archivo de material y transcripción en varios idiomas está siendo empleada por los periodistas de Verifica RTVE con resultados muy satisfactorios (Díaz-Merry, B., conversación personal, 22-11-2023). De hecho, se ha empleado ya para la verificación de los diálogos aparecidos en vídeos de larga duración de la guerra de Ucrania y del asalto al poder de los talibanes en Afganistán, con trascripciones y traducciones del persa o el pastún al español y al inglés.

La segunda de las herramientas desarrolladas dentro del proyecto IVERES se dirige a detectar audios falsos a través de un modelo de inteligencia artificial basado en el adiestramiento de redes neuronales con una base de datos de voces. La herramienta está siendo desarrollada desde la Universidad de Granada, y detecta voces generadas con inteligencia artificial a través de tecnologías de inteligencia artificial. Aunque aún no está disponible para el uso cotidiano de los periodistas, éstos pueden efectuar consultas puntuales al equipo de desarrolladores de la herramienta.

Lo mismo sucede en la herramienta de detección de vídeos falsos y ultrafalsos (*deepfakes*), que se está desarrollando desde la Universidad Politécnica de Barcelona. Los periodistas exponen casos de uso y el equipo de desarrolladores va probando la fiabilidad de la herramienta, aunque aún no se encuentra disponible. Precisamente es a la hora de verificar vídeos con voces e imágenes falsificadas donde los profesionales encuentran actualmente mayores dificultades, por lo que hay grandes expectativas en estas dos herramientas (Díaz-Merry, B., conversación personal 22-11-2023).

Mientras llegan estas dos últimas herramientas, en Verifica RTVE se utiliza el análisis de vídeos frame a frame, y herramientas de libre acceso para verificadores profesionales como INVID We Verify, desarrollada por la agencia de noticias France Press (AFP). Además, en el análisis de vídeos se cuenta con herramientas de libre acceso impulsadas por inteligencia artificial que facilitan las búsquedas inversas desde motores de búsqueda como Google imágenes, Yandex, o Bing, entre otros. Los motores de búsqueda tradicionales efectúan consultas entre textos o entre textos e imágenes. La búsqueda inversa de imágenes es una técnica que va más allá, y que se utiliza para encontrar imágenes idénticas o similares basándose en una imagen de consulta determinada. Esta técnica es utilizada habitualmente

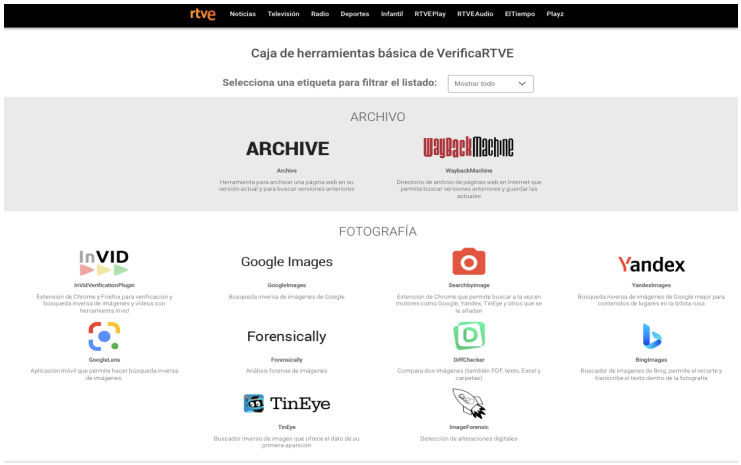
en motores de búsqueda de imágenes que cuentan con amplias bases de datos de imágenes subidas a la web (Nandini et al., 2022).

En Verifica RTVE se ha utilizado también Microsoft Azure en lo que respecta a herramientas de análisis de vídeo, si bien está obteniendo un peor rendimiento y se está empezando a sustituir por otras herramientas.

Finalmente, junto con el apoyo de todas estas herramientas, Díaz-Merry subraya que en la verificación de *deepfakes*, los profesionales apuestan sobre todo por el análisis humano frame a frame, que todavía no puede ser sustituido por herramientas de IA.

La mayor proliferación de vídeos falsos y *deepfakes* se produce en contextos de emergencia como crisis sanitarias, guerras o catástrofes, aunque los verificadores denuncian que la principal tendencia a nivel internacional es que estos vídeos se fabriquen para destruir la reputación de mujeres mediante contenidos sexuales. También se elaboran narrativas propagandísticas desde países en conflicto, como ha sucedido en la guerra de Ucrania y en la de Israel y Palestina, donde se emplean en ocasiones imágenes falsas elaboradas con inteligencia artificial. Los profesionales de Verifica RTVE han hallado ejemplos de contenidos audiovisuales desinformativos en todos esos escenarios. Estos profesionales son conscientes de la necesidad de formarse permanentemente en el uso de nuevas herramientas y de inteligencia artificial, para poder afrontar los desafíos que la propia IA está suscitando en forma de contenidos desinformativos.

Figura 3. Caja de herramientas básica de Verifica RTVE



Fuente: www.rtve.es/noticias/verificartve

5. Discusión y conclusiones

En RTVE la IA se está implementado con una visión global y estratégica y sus responsables pretenden que se infiltre en poco tiempo en todas las áreas de la compañía (Vila, P., conversación personal, 22-22-2023), afectando a un abanico de tareas que incluyen optimizar operaciones, analizar datos y generar nuevos contenidos, en línea con lo afirmado por autores como Sančanin y Penjišević, 2022. La automatización de procesos y la interacción con las audiencias mediante el entrenamiento de algoritmos está mejorando ya la eficiencia de la compañía, de acuerdo a las afirmaciones de Al Husseiny (2023).

En el caso concreto de RTVE se usan herramientas de IA para el análisis de contenido y la creación de contenidos nuevos, como textos, imágenes, voces sintéticas y avatares. En este sentido se emplean herramientas como Lexica, ChatGPT, Dall-E, HeyGen, Studio D-ID, Eleven Labs, Runway y Stable Diffusion, así como las funcionalidades de Adobe. El elenco de estas herramientas, sin embargo, se incrementa a medida que se van desarrollando diferentes proyectos.

RTVE tiene como misión “ofrecer y garantizar el servicio público de radio y televisión de titularidad del Estado” (www.rtve.es/corporacion/quienes-somos), y, manteniéndose fiel a este mandato, debe velar por la correcta transmisión de la información, una información veraz a la que tienen derecho todos los ciudadanos españoles, según recoge la Constitución española en su artículo 20. Es en este marco donde cobra sentido la puesta en marcha de unidades como Verifica RTVE, dedicada específicamente a combatir los contenidos desinformativos y a impedir que se difundan por error dentro de la programación de la radiotelevisión pública.

Para hacer frente al contenido desinformativo, los periodistas de RTVE emplean una combinación de metodologías tradicionales y digitales, dependiendo en buena medida de herramientas gratuitas, unidas a otras adquiridas por la propia corporación, en coincidencia con el estudio elaborado por Haidar (2023).

Para verificar la información, en RTVE se emplean herramientas tradicionales como el análisis frame a frame, junto con un buen número de otras herramientas digitales gratuitas y de pago, muchas de ellas basadas en tecnologías de IA. Tal es el caso de las tecnologías de búsqueda inversa de imágenes que facilitan varios motores de búsqueda, y que se ofrecen de manera gratuita dentro de la caja de herramientas de verificación habilitada en la web de Verifica RTVE.

Si detectar vídeos falsos puede resultar una tarea compleja, aún lo es más la identificación de vídeos ultrafalsos o *deepfakes*, que en RTVE se afronta como un gran desafío debido a la creciente complejidad de las técnicas de manipulación facial, en línea con lo afirmado por Al-Khazraji et al. (2023). Especialmente compleja es la verificación de vídeos falsos compartidos a través de Whatsapp y de *deepfakes* donde no existen voces originales que puedan ser contrastada con las falsificadas (Díaz-Merry, B., conversación personal, 22-22-2023).

En tal sentido, RTVE trabaja en el desarrollo de herramientas propias basadas en IA en colaboración con varias universidades, en el seno del proyecto IVERES, donde se encuentra en fase de pruebas una nueva herramienta de verificación de audios falsos -mediante al adiestramiento a partir de bases de datos de voces- y otra para la detección de vídeos falsos y ultrafalsos.

Las nuevas herramientas pretenden lograr una mayor efectividad en la detección y desactivación de estos contenidos, permitiendo a RTVE ejercer el papel de garante de la veracidad de la información, y garantizar la calidad y confiabilidad de los contenidos emitidos por la radiotelevisión, con el consiguiente incremento de la confianza de los ciudadanos.

6. Agradecimientos

El presente texto nace en el marco del proyecto de Investigación Inteligencia Artificial y Nuevas Fronteras en la Comunicación de la Universidad Internacional de la Empresa (UNIE) Madrid, España.

Especial agradecimiento a los responsables de RTVE por su colaboración y generosidad en brindarnos la información.

Referencias

Al Husseiny, F. (2023). The Rising Trend of Artificial Intelligence in Social Media: Applications, Challenges, and Opportunities. In S. Kaddoura (Ed.), *Handbook of Research on AI Methods and Applications in Computer Engineering* (pp. 42-61). IGI Global. https://doi.org/10.4018/978-1-6684-6937-8.ch003

Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I. y Mishkhal, I. A. (2023). Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications. *The Eurasia Proceedings of Science Technology Engineering and Mathematics, 23*, 429-441. https://doi.org/10.55549/epstem.1371792

Aramburú, L. G., López, I. y López, A. (2023) Inteligencia artificial en RTVE al servicio de la España vacía. Proyecto de cobertura informativa con redacción automatizada para las elecciones municipales de 2023. *Revista Latina de Comunicación Social*, *81*, 1–16. https://doi.org/10.4185/rlcs-2023-1550

Bañuelos, J. (2022). Evolución del Deepfake: campos semánticos y géneros discursivos (2017-2021*). Revista ICONO 14. Revista Científica De Comunicación Y Tecnologías Emergentes, 20*(1). https://doi.org/10.7195/ri14.v20i1.1773

Bouter, M. D. L. D., Pardo, J. L., Geradts, Z., y Worring, M. (2023). ProtoExplorer: Interpretable Forensic Analysis of Deepfake Videos using Prototype Exploration and Refinement. *arXiv (Cornell University)* https://doi.org/10.48550/arXiv.2309.11155

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., y Li, J. (2020). Exploring the Role of Visual Content in Fake News Detection. In K. Shu, S. Wang, D. Lee y H. Liu (Eds.). *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, 141-161. Springer International Publishihaidng. https://doi.org/10.48550/arXiv.2003.05096

Cerdán, V. y Padilla, G. (2019). Historia del fake audiovisual: deepfake y la mujer en un imaginario falsificado y perverso. *Historia y comunicación social, 24*(2), 505-520. https://dx.doi.org/10.5209/hics.66293

De Lima-Santos, M.-F., and Wilson C. (2022). Artificial Intelligence in News Media: Current Perceptions and Future Outlook. *Journalism and Media*, *3*(1), 13-26-. https://doi.org/10.3390/journalmedia3010002

Fernández, A. (2017). Relatos híbridos: El papel de la narratividad en la visualización de información interactiva [Tesis doctoral, Universidad Europea]. Repositorio Abacus https://193.147.239.238/handle/11268/6981

Fernández, A., Revilla, A. y Andaluz, L. (2020). Análisis de la caracterización discursiva de los relatos migratorios en Twitter. El caso Aquarius. Revista Latina de Comunicación Social, (77), 1-18. https://doi.org/10.4185/RLCS-2020-1446

Guarnera, L., y Battiato, S. (2023). An Overview of Deepfake Technologies: from Creation to Detection in Forensics.

Haidar, H. (2023). Using artificial intelligence to verify media content on the Internet. A survey study of journalists working in Iraqi media institutions. *International Journal of Media Studies and Communication Sciences*. https://doi.org/10.36772/arid.aijmscs.2023.485

Hameleers, M., Powell, T. E., Van Der Meer, T. G., y Bos, L. (2020). A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication, 37*(2), 281-301. https://doi.org/10.1080/10584609.2019.1674979

Haseena, S., Saroja, S., Nivetha, A. (2023). TVN: Detect Deepfakes Images using Texture Variation Network. *Inteligencia artificial, 26*(72), 1-14. https://doi.org/10.4114/intartif.vol26iss72pp1-14

Jankowicz, N., Hunchak, J., y Pavliuc, A., Davies, C., Pierson, S., y Kaufmann, Z. (2021) Malign Creativity: How Gender, Sex and Lies Are Weaponized Against Women Online, Washington, D.C.: Wilson Center. https://www.wilsoncenter.org/publication/malign-creativity-how-gender-sex-and-lies-are-weaponized-against-women-online

Jerónimo, P., y Esparza, M. S. (2022). Disinformation at a Local Level: An Emerging Discussion. *Publications, 10*(2), 15. https://doi.org/10.3390/publications10020015

Kalinová, E. (2022). Usage of artificial intelligence on social media in europe. *Ad Alta, 12*(2), 330-333. https://doi.org/10.33543/1202330333

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society, 1*(3), 138-147. https://doi.org/10.1109/tts.2020.3001312

Kavanagh, J. y Rich, M. D. (2018). *Truth decay: An initial exploration of the diminishing role of facts and analysis in American public life*. Rand Corporation.

Liz-López, H., Keita M. , Taleb-Ahmed, A. Abdenour H. , Huertas-Tato, J., y Camacho D. (2023). Generación y detección de contenidos audiovisuales multimodales manipulados: Avances, tendencias y desafíos abiertos. *Fusión de Información*, pp.102-103.

Martin-Rodriguez, F., Garcia-Mojon, R. y Fernandez-Barciela, M. (2023). Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks. *Sensors, 23*(22). http://dx.doi.org/10.3390/s23229037.

Matthews, T. (2023). Deepfakes, fake barns, and knowledge from videos. *Synthese*, *201*(2). https://doi.org/10.1007/s11229-022-04033-x

Nandini S, Akshay B G, Brunda A N, Chandana A M, y Divyashree S R. (2022). Advanced reverse image search and profile creation using machine learning. *International Journal of Advanced Research in Science, Communication and Technology*, 586–589. https://doi.org/10.48175/ijarsct-5417

Olmo, J. y Romero, A. (2019). Desinformación: Concepto y perspectivas. Análisis del Real Instituto Elcano (ARI), (41). https://www.realinstitutoelcano.org/analisis/desinformacion-concepto-y-perspectivas/

Palella, S y Martins, F. (2017). *Metodología de la investigación cuantitativa*. FEDEUPEL

Peña-Fernández, S., Meso-Ayerdi, K., Larrondo-Ureta, A., y Díaz-Noci, J. (2023). Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media. el Profesional de la Información. https://doi.org/10.3145/epi.2023.mar.27

Pineda, A. (2004). Más allá de la historia: aproximación a los elementos teóricos de la propaganda de guerra. En A. Pena (Ed.), *Comunicación y guerra en la historia* (pp. 807-823). Santiago de Compostela: Tórculo. http://hdl.handle.net/11441/64448

Porlezza, C. (2023). Promoting responsible AI: A European perspective on the governance of artificial intelligence in media and journalism. *Communications, 48*(3), 370-394. https://doi.org/10.1515/commun-2022-0091

Rashmi, C., Bhargavi, V., Samhitha, S., Anjana, Y., y Saivaishnavi, V. (2023). Fake detect: a deep learning ensemble model for fake news detection (ml). Turkish Journal of Computer and Mathematics Education (TURCOMAT), 14(03), 684-688.

Sadiku, M. N. O., Ashaolu, T. J., Ajayi-Majebi, A., y Musa, S. M. (2021). Artificial Intelligence in Social Media. International Journal Of Scientific Advances, 2(1). https://doi.org/10.51542/ijscia.v2i1.4

Sančanin, B., y Penjišević, A. (2022). Use of artificial intelligence for the generation of media content. Social Informatics Journal, 1(1), 1-7. https://doi.org/10.58898/sij.v1i1.01-07

Shilpa, B., Kamath, A., Bhat, H., y Sathwik A M. (2023). Unmasking deepfakes: Using Resnext and LSTM to detect deepfake videos. *International Journal of Advanced Research in Science, Communication and Technology*, 524–528. https://doi.org/10.48175/ijarsct-8639

Sohrawardi, S., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R. y Wright, M. (2019). Póster: Hacia una detección sólida de deepfakes en mundo abierto. Actas de la Conferencia ACM SIGSAC de 2019 sobre seguridad informática y de las comunicaciones. https://doi.org/10.1145/3319535.3363269

Stefanov, K., Paliwal, B., y Dhall, A. (2022). Visual Representations of Physiological Signals for Fake Video Detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.08380

Timothy, K., y Shih. A. (2011). Video Forgery. *2011 14th International Conference on Network-Based Information Systems*.

Vedamurthy, H. K., Ravi, y Gururaj. (2022). A reliable solution to detect deepfakes using Deep Learning. *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*.