R&I

IN3
Internet
Interdisciplinary
Institute

research.uoc.edu

OKAYAMA UNIVERSITY

Warsaw University of Technology

CYBERCAT

Arquitectura para la Detección de Noticias Falsas Basada en *Watermarking* y *Machine Learning*

EIG CONCERT JAPAN
Connecting and Coordinating European Research and Technology Development with Japan

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

JST Japan Science and Technology Agency

The National Centre for Research and Development

XVII Reunión Española de Criptología y Seguridad de la Información (RECSI 2022)
Santander, octubre de 2022

Victor Garcia-Font, Tanya Koohpayeharaghi, **David Megías,** Helena Rifà, Julián Salas, Jordi Serra-Ruiz
Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya

Índice

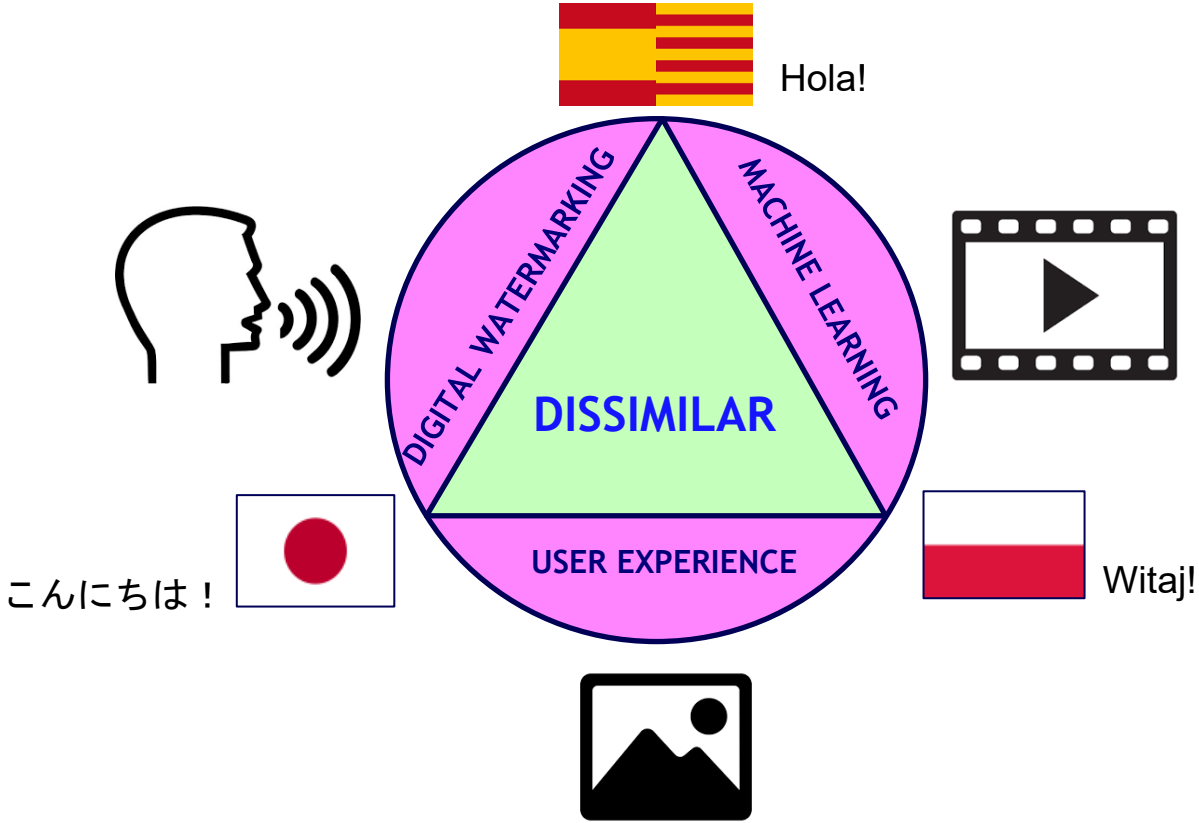
1
Introducción

1. Introducción

- Los líderes de opinión en línea tienen una gran influencia en las comunidades en línea, ¡a veces basándose en noticias falsas!
- La difusión de rumores o de noticias falsas se han visto reforzadas con la creciente relevancia de las redes sociales online y la popularización de la cultura participativa.
- Las marcas de agua digitales se reconocen como una técnica prometedora desarrollada para abordar los problemas de protección de derechos de autor, autenticación de contenidos o detección de manipulaciones, entre otros.

1. Introducción

- ✓ **Interdisciplinario**
- ✓ **Internacional e intercultural**
- ✓ **Plurilingüe**
- ✓ **Múltiples formatos multimedia**



2

Tecnologías

2. Tecnologías – *Watermarking* digital

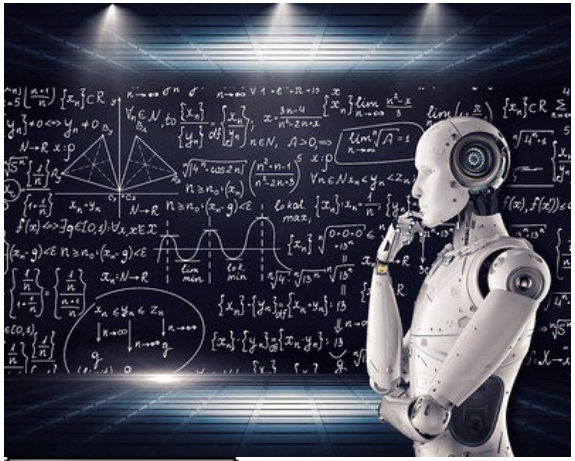
Luis (liamngls)

- Entre las aplicaciones del *watermarking* digital, la prueba de propiedad / identificación del propietario, el seguimiento de transacciones y la detección de manipulaciones son las más relevantes.
- Combinación de diferentes enfoques: marcas de agua robustas para la prueba de propiedad y el seguimiento de transacciones, y marcas de agua frágiles o semifrágiles para la detección y localización de manipulaciones → Escenario novedoso.
- Diferentes tipos de marcas de agua → diversidad de requisitos y mayor complejidad.
- Diferentes tipos de contenido multimedia (imagen, voz y vídeo).

2. Tecnologías – Técnicas forenses para multimedia

- Tecnologías destinadas a revelar la historia de los contenidos: identificación del dispositivo de adquisición, validación de la integridad de los contenidos, recuperación de información de las señales en los contenidos.
- Un dispositivo de adquisición deja huellas específicas debido a sus características intrínsecas: distorsión causada por el hardware.
- Las operaciones de manipulación de contenidos multimedia dejan distorsiones causadas por la operación de procesamiento de señales en el software: distorsión causada por el software.
- Con la ayuda de técnicas de aprendizaje profundo, podemos detectar tales distorsiones y clasificar rastros de edición maliciosos en contenidos multimedia.

2. Tecnologías: detección basada en aprendizaje automático



Mike MacKenzie

https://www.vpnsrus.com

- Clasificación de contenidos falsos → extraer características de un contenido objetivo y un algoritmo de ML calcula una métrica para decidir si las características pertenecen a una clase positiva o negativa.
- Dos pasos: extracción de características y selección de características.
- El uso de aprendizaje profundo permite la extracción automática de características.
- Las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN) son la base de muchas técnicas.
- Se han propuesto varias soluciones para la detección de rostros y voz generados artificialmente.

3

Metodología y arquitectura

3. Metodología

- **El proyecto se divide en tres fases diferentes:**

  - **Fase 1: Diseño e implementación de herramientas de *watermarking***

    Selección e implementación de las soluciones de marca de agua para crear un sistema de prueba de concepto que será evaluada empíricamente.

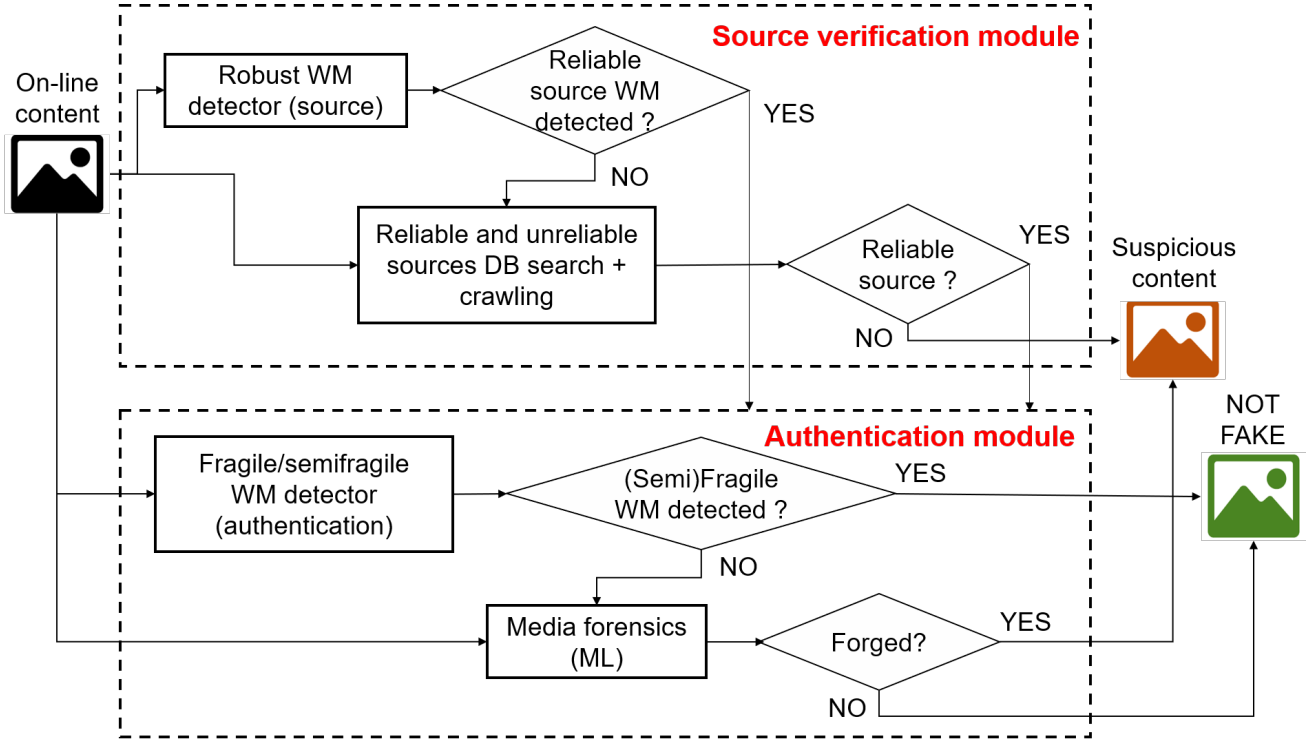
  - **Fase 2: Diseño e implementación de herramientas de detección**

    Diseño e implementación de Redes Neuronales Profundas (DNN) para la identificación de contenidos falsos.

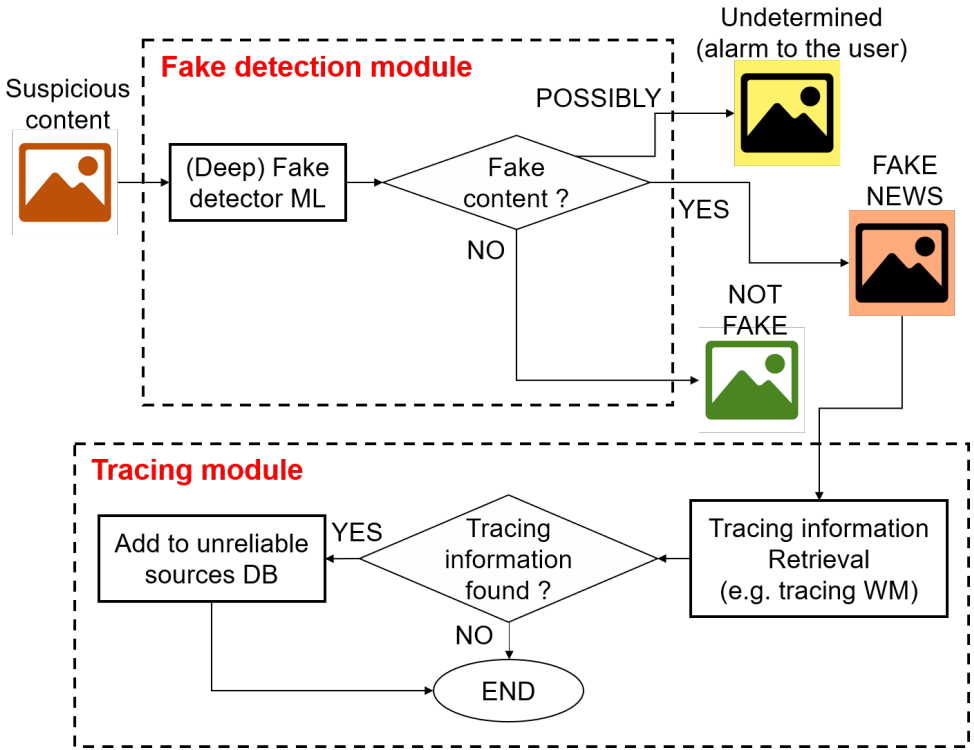
  - **Fase 3: Estudio de la experiencia del usuario**

    ***Focus groups* para una mejor comprensión de los factores culturales involucrados en la difusión de noticias falsas y pruebas de evaluación heurística y usabilidad que combinan datos cualitativos y cuantitativos para el prototipado de la herramienta.**

3. Arquitectura



3. Arquitectura



4

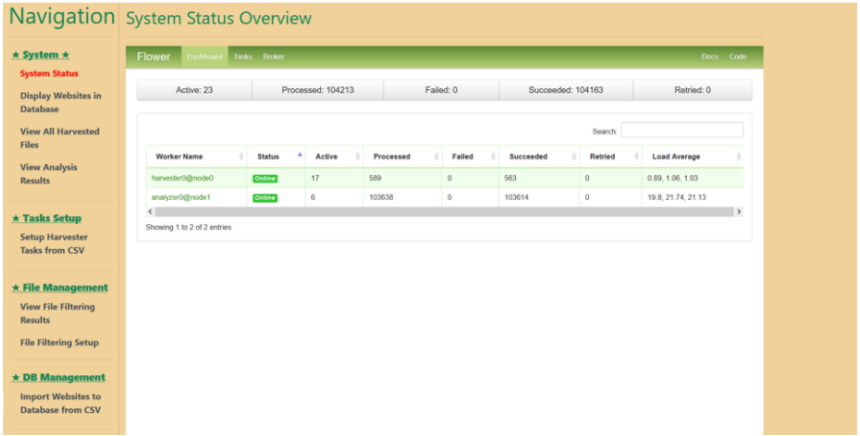
Retos de investigación

4. Retos de investigación

Distributed, Automatized Web Page Harvesting Platform for Large-scale Analysis of Digital Media

The developed harvesting platform is structured as a modular system designed to perform **two main functions**:

- Web page harvesting to obtain digital multimedia content, and
- Analyzing the downloaded content with various analytical tools to detect fake news and other anomalies



4. Retos de investigación

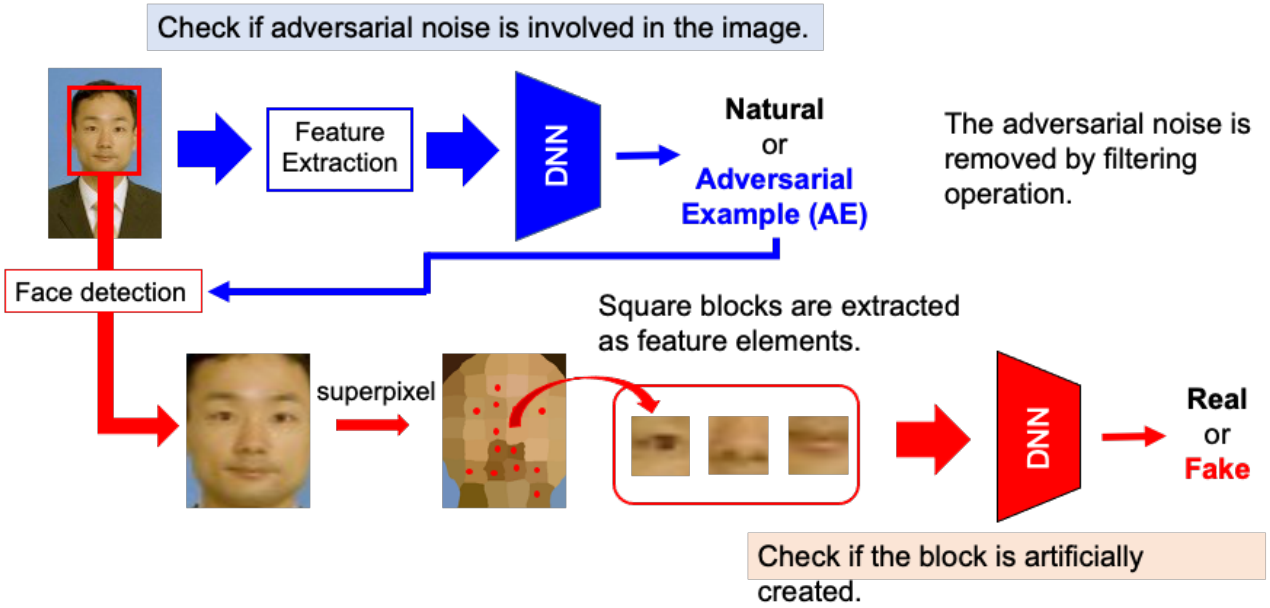
Implementación y combinación de herramientas de *watermarking* para tres tareas diferentes:

1. **Verificación de contenido legítimo proveniente de fuentes reconocidas**
2. **Detección de manipulación maliciosa e intencionada de contenidos**
3. **Trazabilidad de contenidos detectados como falsos**

4. Retos de investigación

Detection of Fake Contents

Two-stage classification architecture for detecting fake contents.



5

Resumen y trabajo futuro

5. Resumen y trabajo futuro

- **DISSIMILAR es un proyecto internacional en curso llevado a cabo por tres socios de España, Japón y Polonia.**
- **El objetivo principal es desarrollar herramientas, basadas en marcas de agua digitales y ML, que permitan la detección de noticias falsas en las plataformas de redes sociales.**
- **Se está realizando un estudio de experiencia del usuario para proporcionar implicaciones para el diseño, la implementación, la integración y la evaluación de las herramientas, sobre la base de las experiencias de diversos usuarios potenciales.**
- **También queremos atraer más investigación de diferentes comunidades académicas para afrontar el desafío de reducir el impacto y la redistribución de las noticias falsas.**

5. Resumen y trabajo futuro

- Proponemos una estrategia interdisciplinaria para abordar el problema que incluye ocultación de datos, ML, análisis forense multimedia y tener en cuenta los factores sociales y culturales.

- Se espera obtener un prototipo que combine todas estas herramientas y conocimientos sobre el comportamiento de los usuarios y que sea más exitoso que las soluciones parciales por sí solas.

¡¡Muchas gracias!!

http://dissimilar.ii.pw.edu.pl/