Architecture of a fake news detection system combining digital watermarking, signal processing, and machine learning

David Megías[1,2], Minoru Kuribayashi[3], Andrea Rosales[1], Krzysztof Cabaj[4], and Wojciech Mazurczyk[4*]

[1]Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain
{dmegias, arosales}@uoc.edu

[2]CYBERCAT-Center for Cybersecurity Research of Catalonia

[3]Okayama University, Okayama, Japan
kminoru@okayama-u.ac.jp

[4]Warsaw University of Technology, Warsaw, Poland
{krzysztof.cabaj, wojciech.mazurczyk}@pw.edu.pl

Received: November 30, 2021; Accepted: February 10, 2022; Published: March 31, 2022

Abstract

In today's world, the ease of creation and distribution of fake news is becoming an increasing threat for individuals, companies, and institutions alike. Content spread over the Internet is able to create an "alternative" reality and false accusations cannot be easily removed by later issued apologies as it typically takes several years to unpick the labels pinned on by spreading disinformation. Currently, the main facilitators of fake news distribution are social media networks, where a large volume of digital media content is generated and exchanged every day. In this "flood" of information, it is quite effortless to manipulate the content to impact its consumers. That is why developing effective countermeasures is of prime importance. Considering the above, in this paper, we propose and describe an architecture of the fake news detection system that is being developed within an ongoing Detection of fake newS on SocIal MedIa pLAtfoRms (DISSIMILAR) project. It is designed for the protection of digital media content, i.e., images, video, and audio, and to fulfill its goals, it combines digital watermarking, signal processing, and machine learning techniques.

Keywords: Fake news, digital watermarking, machine learning, signal processing, user experience study.

1 Introduction

Fake news and hoaxes, which allow spreading various types of disinformation to influence certain groups of people or whole societies, have been present in the history of humankind even before the advent of the Internet. In general, fake news is typically considered as a kind of yellow journalism in which fake news encapsulates pieces of legitimate news that may be hoaxes. As already hinted, such actions are typically performed to impose certain ideas so that they are considered publicly as legitimate ones.

In recent years, the accelerated adoption of social media platforms enabled rapid information sharing that has never been present in human history before. Using social media networks, their users are able

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 13(1):33-55, Mar. 2022
DOI: *10.22667/JOWUA.2022.03.31.033*

[*]Corresponding author: Nowowiejska 15/19, 00-665 Warsaw, Poland, Tel: +48 22 234 77 11, Web: http://mazurczyk.com

to create and share more information than ever before. Unfortunately, some of this news are deliberately deceitful. That is why, in the current digitalized world, the main channels to spread disinformation are social media platforms and other types of online media. The fake data include not only text information but increasingly also manipulated digital images, videos, or audio files.

The spread of fake news across social media platforms has already impacted events in real, non-digital life. For example, in 2016, during the US presidential election, various kinds of fake news about the candidates were widely disseminated in social networks. This, as it has been reported in [1], had a potentially significant effect on the real election process as it was estimated that online social networks account for more than 41.8% of the fake news data traffic in the election. This was far more effective and widespread than traditional channels (i.e, TV, radio, or printed media). Another recent example is related to the ongoing SARS-CoV-2 coronavirus pandemic. The disinformation campaigns related to the virus itself, its severity, origin, potential ways of infection, treatment, and finally vaccination have started spreading faster than the virus itself. The latter example proves that fake news causes real harm, resulting in deaths of real people throughout the world.

It must be pointed out that there exist two distinct types of fake news. First, some fake news are created from contents that are legitimate in origin but have been manipulated in malicious ways, e.g., by replacing the audio of a video clip or even creating a deep fake from an authentic video. Second, we can find fake news that are created from scratch without manipulating a legitimate original content. As discussed below, the project aims at addressing both types of fake news.

To address above-mentioned issues, in June 2021, we have started the Detection of fake newS on SocIal MedIa pLAtfoRms (DISSIMILAR) project whose aim is to equip the content creators with solutions that will be able to add watermarks to the content they create and make any modification easily detectable. Additionally, this would enable online social media users to apply tools based on state-of-the-art signal processing and machine learning (ML) methods to detect fake content. To achieve this aim, in DISSIMILAR, we will create models to detect fake digital media content, focusing on the distortions created by the signal processing operations and recording devices. The combination of watermarking and ML-based detection tools will empower users to discriminate between original and fake multimedia content without the need for assessment and control from a centralized service.

Apart from the technical advances, the DISSIMILAR project will carry out a cross-cultural user experience design approach [2] to define implications for the design [3] in all stages of the process. To design and develop tools that are usable, useful, and appealing, this project will conduct a comprehensive user experience study. Finally, it is worth noting that the collaboration of three partners from Japan, Poland, and Spain is fundamental not only to complement their expertise and technical background but also to build on participants from different regions, cultural backgrounds, and life trajectories in the user experience study that allows to take into account the values of potential final users in different contexts.

This paper is an extended version of our conference publication [4]. It must be emphasized that in contrast to this previous work, the novel scientific contribution introduced in this article can be summarized as follows:

- We present in detail the architecture of the proposed fake news detection system, focusing mainly on how different techniques, i.e., signal processing, digital watermarking, and machine learning, are interacting with each other to form a complete system,
- We outline the initial project's test-bed, which contains dedicated custom software that is used for acquisition of real-world multimedia content from the Internet. This would allow evaluating various fake news detection methods known from state-of-the-art publications and developed within the course of the project.

With this work, we expect to raise awareness on the disinformation challenge in the data hiding and machine learning research communities. We also create the architecture of the DISSIMILAR system in an elastic and easily extensible manner. Thus, we hope to attract new research that can be integrated into the proposed solution in the future. Therefore, we contribute to limiting the influence of fake news with a decentralized solution that does not lead to censorship or biased and interest-driven decisions. The project will provide a first set of technologies, combining digital watermarking and machine learning solutions, to be integrated with social media platforms for fake news detection. The prototype will allow adding and replacing different components, such as digital watermarking algorithms or machine learning models. Hence, we expect to provide a mechanism to integrate technologies from other contributors in the platform.

The remainder of the paper is structured as follows. In Section 2, we present the works that are most related to the topic of this paper. Then, in Section 3, we explain the fundamentals needed to understand the concept of the proposed framework. Next, Section 4 presents the overall design of the DISSIMILAR project. In Section 5, the main research phases of the project are outlined. In Section 6 the envisioned overall project architecture is described as well as the detection process is characterized with the most important steps and modules, while in Section 7 the planned evaluation platform is presented. Then, Section 8 showcases the expected impact of the project. Finally, Section 9 concludes our work and presents some future research directions.

2 Related work

Mass-self-communication (MSC) refers to horizontal networks of communication where users become both senders and receivers of messages, e.g., social network sites (SNS) [5]. In MSC, stories are told differently, as different voices could be involved, including users able to participate in the expansion of the story in different ways [6]. “A mix of top-down and bottom-up forces determine how the material is shared in far more participatory (and messier) ways” [6, p.1] than in hierarchical media. With digital media, the idea of prosumers is reinforced [7], as users become producers and consumers able to circulate and recreate contents. People consume, share, reframe, mix, and create media beyond the paradigms of one-to–to-one communication or one-to–to-many spectators. This shift from distribution to circulation builds on the participatory culture that flourished with digital media, which, at the same time, gives new opportunities to the creation and dissemination of fake news.

In the early 2000s, with the diversification of sources providing online news, and the use of social network sites to filter news, an early alarm prompted the risk of some individuals getting trapped in “filter bubbles”, or “echo chambers”, where they mostly get information that confirms their intuition [8]. Online opinion leaders have an influence on online communities, sometimes building on fake news [9]. Moreover, the use of big data, and artificial intelligence allows influencing individuals building on their predicted ideology, but also their predicted fears and phobias, in a “information psychological war” [10]. Thus, the spread of rumors [11], or the dissemination of fake news [12], were reinforced with the increasing relevance of SNS and the popularization of the participatory culture [13].

The spread of fake news has demanded news media a more significant effort to show they stand for integrity. In this context, different projects emerged. On the one hand, fact check projects devoted to unmasking hoaxes have emerged in several countries, e.g., FactCheck.org (USA), maldita.es (Spain), or Demagog.org.pl (Poland), or FactCheck Initiative (Japan) and have been widely adopted by consumers. Such websites usually show the metadata of multimedia contents to argue the originality of the media. However, metadata can be easily changed and does not show what has been changed in the content, so they require further checks.

On the other hand, the Trustproject [14] provides a protocol that includes eight trust indicators to ”To amplify journalism’s commitment to transparency, accuracy, inclusion, and fairness so that the public can make informed news choices "[14]. The protocols have been adopted by hundreds of news sites worldwide, including prominent media companies such as BBC, South China Morning Post, and Bay Area News Group. One of the indicators refers to the ability of the consumer to identify the journalist's expertise. Who made this. While such projects have been well received, they lack technological tools to automate the procedures.

In this context, digital watermarking is recognized as a promising technique developed to address the problems of copyright protection, content authentication, tamper detection, and others [15]. In some watermarking applications, a unique fingerprint identifying the recipient of a multimedia content is embedded in each individual copy of the distributed content. This application acts as a deterrent to illegal redistribution by enabling the owner of the content to trace the source of the redistributed copy [16, 17].

Another promising venue where the application of digital watermarking techniques could be beneficial is fake news identification and tracing. Such a concept has not been so far researched in the existing literature, and thus it can be considered as novel and interesting. Although there have been some attempts to counter deep fakes [18] or fake news in images, they have not been proposed and analyzed for other types of digital content, and they have never been applied in a more complete system integrated with social media platforms. Thus, we can conclude that the proposed approach has innovative potential. Regarding detection techniques, typically existing works for detecting fake videos are centered on finding imperceptible characteristics that appear in the forged videos. For example, the method proposed in [19] is based on the detection of eye blinking, which is a physiological signal that is not well presented in the synthesized fake videos. The method in [20] visualized the CNN layers and filters and discovered that the eyes and mouth play a paramount role in the detection of faces forged with the deepfake software tools [21, 22].

With the understanding of fake video problems, the need for creating automated detection methods not only in academia, but also in industrial environments is of utmost importance. While digital forensics experts have developed different individual methods for some small instances, the hundreds of thousands of videos uploaded to the Internet or social media platforms have a variety of scale and qualities. To accelerate advancements in the detection of fake media, the DeepFake Detection Challenge (DFDC) data set [23] was constructed and publicly released by Amazon Web Services (AWS), Facebook, Microsoft, the Partnership on AI's Media Integrity Steering Committee and academics. The goal of the challenge is to stimulate researchers around the world to build innovative new technologies that can help to detect deepfakes and manipulated media.

In contrast to the existing work presented above, in the DISSIMILAR project, we aim to develop models to detect fake digital media content by focusing on the unnatural signals created by the signal processing operations and recording devices. The signals are intentionally inserted as watermarks, whereas the traces of modification and editing are left behind when the fake contents are created. The combination of watermarking and ML-based detection tools will allow users to easily discriminate between original and fake content without the need for assessment and control from a centralized service.

3 Background

This section introduces a brief overview of the techniques that will be used throughout the project, including digital watermarking, machine learning for the detection of fake news, and user experience (UX).

3.1 Digital watermarking

Digital watermarking –a branch of data hiding [24]– is a collection of techniques consisting of embedding data, referred to as a mark or a watermark, into a digital object or carrier, typically maintaining the perceptual quality of the object. The traditional carriers of watermarking are multimedia contents, such as images, audio, or video, but also text and even network protocols. The embedded mark is related to the cover object, and the cover object itself is valuable (often more valuable than the watermark). The watermark can be used, for example, to provide evidence about the copyright holder or the authorized viewers of the content. Traditional applications of digital watermarking include copyright protection, content authentication, broadcast monitoring, transaction tracking, and copy control, among others.

Another well-known branch is steganography, which aims at transferring secret information between two communicating parties. Unlike cryptography, the objective of steganography is not making a piece of information unreadable for those who are not authorized, but it makes the exchange of information itself secret by hiding the communication in an apparently innocuous carrier. In the case of steganography, the cover object is typically considered useless and the item to be protected is the secret message. It is, thus, important to remark that digital watermarking is not a particular form of steganography, but another data hiding branch that shares some common features with steganography but with different properties and applications.

Digital watermarking schemes are often analyzed in terms of five main properties, namely, capacity, robustness, transparency (or imperceptibility), blind or informed detection (or extraction), and security.

Capacity or data payload refers to the amount of information that can be carried by the marked object. Usually, the amount of data is given in bits per unit, where the unit depends on the type of object. Some examples are bits per second and bits per pixel.

Blind or informed detection or extraction refers to the requirement or not of the original (cover) object when carrying out the detection (or extraction) of the hidden watermark. If the cover object is required at the detection/extraction end, then the detector/extractor is referred to as non-blind or informed. The whole scheme is usually called *a non-blind watermarking scheme or an informed watermarking scheme*. When the original cover object is not used by the extraction or detection algorithm, then the detector/extractor is referred to as *blind*, and the watermarking scheme as *blind watermarking*.

Transparency or imperceptibility is related to the perceptual quality of the marked object compared to the cover (original) object. For example, if the marked object is an image, the embedded information will make some pixels of the marked image differ from the corresponding pixels of the cover image. Imperceptibility depends on the amount of "perceptual noise" introduced in the image by the embedding process. Hence, imperceptibility can be defined as the "perceptual similarity between the cover and the marked objects".

Robustness refers to the ability of the watermarking scheme to detect or to extract the embedded watermark when the marked object is transformed using standard signal processing operations, such as filtering, lossy compression, or geometric transformation. A watermarking scheme that can resist signal processing attacks is referred to as *robust watermarking*. On the other hand, some applications require that the watermarks are removed when any or some transformations are applied to the watermarked object. In this case, the term of fragile (no transformations allowed) or semi-fragile watermarking (some transformations allowed) is used.

Security is the ability of a watermarking scheme to resist hostile attacks. In this case, we can distinguish two categories of attacks, namely, attacks against the embedded watermark and attacks aimed at the (secret) keys of the watermarking scheme. Attacks of the former type can be classified in three categories: unauthorized removal, unauthorized embedding, and unauthorized detection (or extraction).

Unauthorized removal refers to the possibility of suppressing or masking the embedded watermark such that it cannot be detected or extracted. As unauthorized embedding (or forgery) is concerned, the purpose of the attacker is to embed a watermark into a work to provide a false authentication of the contents. Both unauthorized removal and embedding are active attacks, but there is a third possibility that can be thought of as passive: unauthorized detection (or extraction).

The second category of security attacks is related to the watermarking keys used in the scheme. Similarly to cryptography, many watermarking schemes require the use of secret (often symmetric) keys to be applied both for embedding and detection or extraction. However, in watermarking schemes, some keys that are not identical to the ones used by the embedder, but close to them, may lead to the successful unauthorized extraction or removal of the watermark. Randomization in some parts of the watermarking algorithms is a common tool to increase security.

In addition, the embedded message or watermark can itself be encrypted before embedding and decryption will be necessary in the receiving end to decipher the embedded data. This combination of watermarking and encryption, each with its particular key (watermarking and crypto/cipher keys, respectively), is common in most data hiding applications for security reasons. Other digital watermarking properties include embedding efficiency, false positive rate, modification and multiple watermarks and computational cost.

Among the applications of digital watermarking that can constitute the basis for the DISSIMILAR project, owner identification/proof of ownership, content authentication/tampering detection and transaction tracking are the most relevant ones. In **owner identification/proof of ownership** applications [25], a watermark can be used to provide contact information about the owner or source of a given work. This idea extends the common textual copyright notices found in these works, since the watermark is invisible and inseparable from the marked object. In the case of proof of ownership, the goal of the watermark is not only to identify the owner, but to prove who the owner of the work is, even in court. In **content authentication/tampering detection or localization** applications [26, 27, 28], the embedded watermarks can also be used to detect whereas the marked work has been tampered by an adversary. In addition to a yes/no detection (tampering detection), some systems also allow specific forgeries to be localized in the content (tampering localization). For this application, fragile or semi-fragile watermarking systems are required, compared to the robust schemes that are used in many other applications. In **transaction tracking/fingerprinting** applications [29, 30, 31, 32], different watermarks (called fingerprints) are embedded in a content that is distributed to different users. Each copy of the content is embedded with a unique fingerprint that is used to identify the user in case he/she decides to redistribute the content (tracing).

The application of digital watermarking in the field of fake news detection will require a combination of different approaches, possibly involving robust watermarks for proof of ownership and transaction tracking, and fragile or semi-fragile watermarks for tampering detection and localization [33]. This is a challenging application scenario, since different types of watermarks have a diversity of requirements and the resulting embedding and detection methods will entail more complexity. In addition, the system must be designed to work with different types of multimedia content (image, audio, and video).

3.2 Multimedia forensics

In the case of steganography, a malicious party may use the technique for secret communications over a public channel without being suspected by possible eavesdroppers. As a countermeasure for steganography, the analysis of hidden messages in multimedia content, known as steganalysis, has been intensively investigated for classifying content with/without hidden messages.

Motivated by the study of steganalysis, the irregularity of multimedia content has been measured from the forensics point of view. Multimedia forensics includes a set of scientific techniques recently proposed for the analysis of multimedia content such as audio, video, and images to recover evidences from them. In particular, such technologies aim at revealing the history of content:

- identification of the acquisition device that produces the data,
- validation of the integrity of the content,
- retrieval of information from the signals involved in the content.

For source identification, it is assumed that an acquisition device leaves specific traces due to its intrinsic characteristics (e.g., sensor noise, lens distortion, and others), which comes from the hardware-oriented distortion. Similarly, tampering operations of multimedia content leave distortions caused by the signal processing operation in software (e.g., lossy compression, filtering, and others), which is called software-oriented distortion. Different processing algorithms may produce identifiable traces, and some inconsistencies of scene characteristics are introduced by tampering. With the assistance of deep learning techniques, we can detect such distortions and classify malicious editing traces in multimedia content. Such multimedia forensics techniques will enable us to detect fake content with a high accuracy.

3.3 ML-based detection

ML algorithms learn the mapping from an input to an output. In the case of classification problems, the algorithm learns the function to separate two (and sometimes more) classes for a given task, which is known as the decision boundary. The decision boundary helps in determining whether a given data point belongs to a positive class or a negative class. For the classification of fake contents, we first extract features from a target content and an ML algorithm calculates a metric whether the feature belongs to a positive or a negative class. Here, we should consider two steps in the ML model: feature extraction and feature selection. In feature extraction, we extract the required features for a given task. On the other hand, in feature selection, we select the important features that improve the performance of an ML model.

Generally, it is time-consuming to construct the feature extracting function manually from an image, and it needs specific knowledge of the subject as well as the domain. By using deep learning techniques, the tuning of such a function can be automatically calculated. Among some branches of deep learning techniques, Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are the basis of many techniques. RNNs capture the sequential information present in the input data, i.e., the dependency between the words in the text, while making predictions. On the other hand, CNNs capture the spatial features from an image. Spatial features refer to the arrangement of pixels and the relationship between them in an image. They allow identifying the object accurately, the location of an object, as well as its relation to other objects in an image.

A deep learning-based technique enables us to create fake contents by swapping the face of a person with the face of another and by synthesizing the movement of the face according to the manipulated audio speech. The facial manipulations can be categorized into four groups: *Entire Face Synthesis*, *Identity Swap*, *Attribute Manipulation*, and *Expression Swap* [22].

As a countermeasure of such fake contents, there are some studies to classify whether faces are real or artificially generated. The first studies in this area focused on the audiovisual artifacts existing in the 1st generation of fake videos. The inconsistencies between lip movements and audio speech are analyzed in [21]. In [34], the detection capability is improved by using the Long Short Term Memory (LSTM) which is based on RNN. Some simple visual aspects such as eye color, missing reflections, and missing details in the eye and teeth areas have been utilized in [35] for the classification of fake contents by using a logistic regression model and a multilayer perceptron (MLP) [36]. A detection system based on both facial expressions and head movements is proposed in [37], which employs a Support Vector Machine (SVM) [38] for the final classification. The change of eye blinking patterns is studied in [39].

From the detected face regions and the surrounding areas, a detection system [19, 40] based on CNN, such as VGG16, ResNet50, ResNet101, and ResNet152, detects the presence of artifacts incurred by the difference of resolution. Approaches based on mesoscopic and steganalysis features are proposed in [20], in which a CNN-based system is intensively investigated. In this work, the detection system based on the XceptionNet architecture provided the best results comparing different alternatives.

4 Project design

This project can be divided in two main areas: (i) the development of technological tools to assist users to distinguish between original and forged content in multimedia publications, and (ii) a case study to include the cultural dimension and to center the development on real users' needs and behavior. In the following subsections, we provide details about the above-mentioned areas. https://www.overleaf.com/project/61768ceece30f

4.1 Watermarking and detection tools

The project will provide a set of technological tools to assist users to identify original and forged content in multimedia documents (i.e., sound, images, video). For that, the DISSIMILAR project will provide two types of tools:

- *Watermarking tools*: In this project, we will design and develop a watermarking-based system to embed information in media files before they are published on social media platforms. Once they are published, it will be possible to easily and automatically verify whether a media document comes from a trusted source or whether it has been altered to create fake news. To achieve this goal, a set of digital watermarking techniques will be applied to embed authentication watermarks, which can be imperceptible or visible (or audible), in the original file and will not be easily removable without affecting the content. To this aim, any applied modification should be easily detectable and automatically identifiable. As a result, the developed system, integrated within online social media platforms, will be able to protect any kind of digital media content (i.e., images, video, audio) and reliably warn users when they receive forged content. Additionally, in many cases, it will be also possible to identify the origin of the fake content propagation. Considering the above discussion, the main objectives associated with this research area are the following: (i) selection of suitable digital watermarking techniques; (ii) design of the architecture of the proposed system; (iii) design and development of the automatic and easy-to–to-use digital content verification mechanism with the user warning feature if fake news is detected; (iv) design and development of the mechanism for identifying the origin of the fake content propagation; and (v) development of the proof-of-concept implementation of the complete system.

  Digital watermarking has some advantages to other detection methods, such as ML solutions. To begin with, it does not require a training set and, hence, it avoids problems like that of overfitting. The presence or absence of a watermark can be enough to discriminate between authentic or forged contents. In addition, the combination of different types of watermarks, e.g., audio, video, robust, and fragile, can be very powerful for the detection of forgeries. Furthermore, watermarking can also be used for data tracing (data provenance) and, thus, it may allow identifying the source of a forged content.

- *Detection tools*: In many cases, fake news are distributed in direct messaging applications out of the traditional online social media. Therefore, embedding information of any type is not always effective. Furthermore, we cannot expect that all media channels and social networks will use the proposed (or any other) watermarking scheme before publishing media files. Hence, it is important to provide tools that can detect whether some media content that has not been previously marked is forged or not. To address this challenge, the project will also provide a tool based on ML to analyze tiny unnatural signals induced by the generation of fake content. To this end, this project will create a tool that mixes two approaches focusing on: (i) distortions caused by signal processing operations, such as ML tools, which can be classified as software-dependent characteristics, and (ii) device-oriented distortions introduced by differences in the recording devices, which can be classified as hardware-dependent characteristics. The main objectives associated with this research area are the following: (i) extraction of feature vectors from suspicious multimedia contents; (ii) design of an artificial neural network (ANN)-based architecture for classification; (iii) collection of datasets for training the proposed system; (iv) development of a compact and efficient implementation; and (v) development of the proof-of-concept implementation of the complete system.

4.2 User experience study

A user experience (UX) study will provide implications for the design, implementation, and integration and evaluation of the tools, based on the experiences of diverse potential users from the global north. The study will be conducted in Japan, Poland and Spain. The three countries are high income countries according to the World Bank [41], they are from Asia and Europe. However, they represent different cultural and historical backgrounds that could influence the consumption of fake news. In addition, fake news are mainly consumed in the national languages. However, Japanese and Polish are mainly restricted to the geographical context, while fake news in Spanish could come from different contexts, which make the three of them as three different and interesting cases to study.

- Identify cultural factors that should be taken into account in the design of the system. This implies analyzing how users interact with news in general and, in particular, in relation to the objectivity, credibility, and accuracy of the information. The study will analyze the relevant features of news that increases or decreases its credibility to subjects. We will also conduct a detailed investigation on how the proposed tools can contribute to: (i) increasing the subject's awareness on fake content and (ii) minimizing the spread of fake information.

- From a human-computer interaction point of view, the study will analyze the best ways to offer tools to the subjects. For instance, we will study which type of watermark is more effective, which information must be embedded in the different types of files, what kind of fake information is detected by the subjects, and so on. Moreover, the study will focus on the best procedures to interact with the proposed tools and the optimal ways to communicate the results obtained from the watermarking and ML analyzers.

- Providing results that are relevant to the three countries participating in this project (i.e., Japan, Poland, and Spain) will allow cultural comparative research. For this purpose, the case study will be adapted to the cultural background of each of the above-mentioned countries.

- Contribute to the iterative prototyping process to design and deploy the proposed tools. This case study is scheduled to start from the beginning of the project to involve the users' point of view and the cultural knowledge gathered around them in the development of the tools as soon as possible.

The main objectives associated with this research area are the following: (i) identify the cultural factors that should be considered in the design of the tools; (ii) take critically informed decisions for the design of the tools; and (iii) evaluate the potential impact of the system.

4.3 Scientific excellence

This project proposes an innovative approach to combat fake news in multimedia content based on three main research activities:

- A comprehensive user experience study to put the user at the center of this research. This study will analyze cultural and behavioral aspects that are essential to build effective tools. The three partners will participate in the study to provide relevant data to adapt the tools and measure their success in three different countries.
- Design of watermarking tools to take a novel approach to automatically identify manipulations in multimedia documents in a complete system that can be integrated with social media platforms.
- Design detection tools based on state-of-the-art ML and signal processing methods to assist users to determine when multimedia documents that do not include watermarks have been manipulated.

The main strength of the DISSIMILAR project is the combination of these three research activities, which will provide more generalizable results than conducting research separately, achieving exploitable outcomes, and an integrated and user-centric prototype.

4.4 Added values of multilateral cooperation

The multilateral cooperation of institutions from three different countries (Japan, Poland, and Spain) is essential for the following reasons:

- The influence of fake news and different disinformation techniques may have a different effect on people from different backgrounds and cultures. For this reason, it is important to undergo a case study of this project involving three partners to analyze the problem at least from a European perspective, including two countries with very different cultural backgrounds (i.e., Poland and Spain), and involving also the Japanese perspective.
- The development of the tools will be carried out with a user-centered design. Users will be involved at an early stage of the project. Multilateral cooperation in this aspect is of utmost importance to take into account the human-computer interaction point of view beyond the national/regional angles of a single partner.
- Measurements of success of the proposed tools will be taken involving all partners and subjects from those three countries.
- From a scientific point of view, multilateral cooperation is essential to cover the three main areas of the DISSIMILAR project. In this regard, the Polish institution has experts in watermarking, Japan has experts in signal processing and ML, and the Spanish partner brings an interdisciplinary team with technical expertise in digital watermarking and social scientists that excel in executing case studies and human-computer interaction analyses.

5 Research phases

To protect users from fake news, the DISSIMILAR framework offers three different technical components, i.e., (i) forensics, (ii) watermarking, and (iii) ML-based detection tools. Forensics tools are aimed to evaluate whether the digital media produced by content producers that should enrich the online social media platform present manipulated content or not. Next, the watermarking tools are responsible for providing the digital “stamp” on the introduced content. Moreover, ML-based detection tools allow users to determine the genuineness of the received content.

As explained in the previous sections, the DISSIMILAR tools would be designed and developed by taking into account the results of the multinational user experience study. Note that all components of the proposed framework are described in Section 4.

In summary, the research activities in the DISSIMILAR project are divided into three phases:

Phase 1: Design and implementation of watermarking tools,

Phase 2: Design and implementation of detection tools, and

Phase 3: User experience case study.

Below, we describe the research methodology that will be followed in each of those project phases.

5.1 Phase 1: Design and implementation of watermarking tools

In this phase of the project, we will use a mixed approach that combines qualitative and quantitative methodologies. In particular, we will use proof-of-concept implementations of the components of the designed digital watermarking system to experimentally evaluate their performance and whether they fulfill the expected requirements. The proof-of-concept implementation of the complete developed system will be subjected to the experimental evaluation as well.

Digital watermarking is a tool for embedding information into multimedia contents, and our scope involves considering its different applications. For instance, receivers of the content may be able to trace its original source of contents, and the detector could use the embedded marks to identify user(s) from the distributed fake contents.

Watermark techniques can be classified into two categories: robust and fragile. Robust watermark is resistant to any modification of the content. While a fragile watermark is vulnerable to any modification that can be easily changed by modifying the content. The latter can be used to detect intentional modifications. For the detection of fake contents, we combine these two watermarking techniques in this project.

5.2 Phase 2: Design and implementation of ML detection tools

In essence, multimedia contents are created by using recording devices such as digital cameras and microphones. Due to the difference of optical and sensor devices, hardware-oriented distortions must be contained in the captured multimedia content. In a forensic case, multimedia contents are crafted by editing and combining multiple resources. This process must cause some distortions, including unnatural signals and noise with different characteristics. We will analyze the unnatural signals involved in fake contents by using both signal processing and ML techniques.

The unnatural signals come from the characteristics of hardware devices and software operations. It will allow us to find the origin of the content as well as the traces of recapturing. Meanwhile, software-oriented distortions mainly come from nonlinear operations, such as rounding and lossy compression. Even if a malicious party creates fake contents by using deep learning techniques, the unnatural signals will help us to classify them as fake.

The consistency of characteristics, both in the time and the frequency domain, is difficult to control in the artificially created contents as well as manipulated ones. The technique for analyzing unnatural signals in multimedia contents has an analogy to the technique of steganalysis - the detection of steganography. The detection of the existence of the hidden message is explored during steganalysis with, e.g., the assistance of ML techniques [42, 43]. Similarly, the distortions introduced during the creation of fake contents can be analyzed using a framework analogous to that of steganalysis. A sophisticated analysis of unnatural signals is investigated to design an efficient Deep Neural Network (DNN)-based classifier considering the characteristics of the distortions.

In the past few years, a significant number of researchers have investigated fake content detectors based on ML combined with conventional signal processing techniques. However, the progress of the detector stimulates other researchers to craft more natural looking fake contents and make them difficult to be classified with the assistance of a Generative Adversarial Network (GAN) architecture. In addition, there are some reports about the jamming attack to the detectors by adding adversarial noise, which is crafted to cause misclassification [44]. The problem of adversarial noise is recognized as one of the threats for ML systems because it will fool DNN-based classifiers without seriously degrading the contents [45]. The robustness against adversarial noise is also required to design a good fake content classifier.

5.3 Phase 3: Case study

We plan to use a cross-cultural user-centered design approach [2] to understand and quantify differences or similarities in the user experience. We will use focus groups in the initial phase of the project for a better understanding of the cultural factors involved in the dissemination of fake news. We will also use heuristic evaluation and usability tests that combine qualitative and quantitative data for the iterative prototyping of the tools and the potential impact of the project. Of course, one of the factors to consider in this phase are linguistic issues, since we will target at least three different languages, namely, Japanese, Polish, and Spanish, in the case study.

6 Envisioned architecture and comparative analysis

As discussed in the previous section, DISSIMILAR will combine different technologies to help users in fake news detection. A description of the architecture of the system is required to understand how those different technologies can be used jointly to (try to) identify fake news. We also present a comparative analysis with other architectures.

6.1 Envisioned architecture

The high-level overview of the envisioned DISSIMILAR framework architecture is illustrated in Fig. 1. In the assumed scenario, we distinguish content producers and content consumers. The former can generate genuine or fake news and feed them into online social media platforms. The latter are typical social media platform users which are targeted by malicious content producers with manipulated information.

Moreover, the proposed detection process is divided into two steps and each step consists of two modules:

1. **Step 1: Source verification and content authentication.** This phase is intended to verify if the source of an on-line content (audio/voice, video or image) is reliable or not, and try to determine if the content has been modified or not.

   A flowchart of this step is displayed in Fig.2. It can be observed that this phase consists of the following two modules:

   (a) **Source verification module.** It tries to verify the source of the content using either a robust watermark detector (the source should have embedded a robust watermark before distributing the content) or (if no watermark is detected) making a search on the Internet to locate the

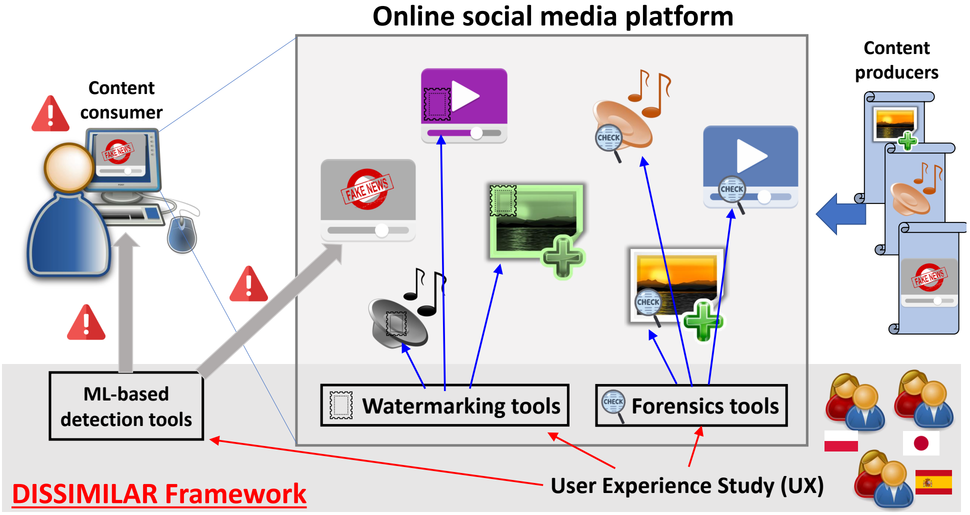

Figure 1: Envisioned DISSIMILAR framework high-level architecture

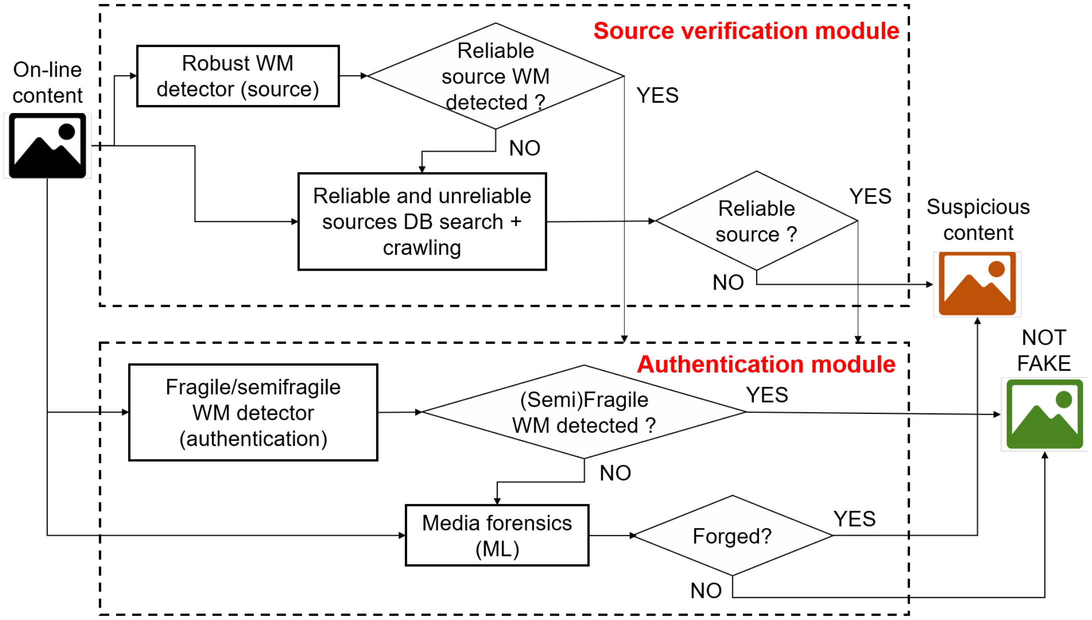
Source verification module
On-line content
Robust WM detector (source)
Reliable source WM detected ?
YES
NO
Reliable and unreliable sources DB search + crawling
Reliable source ?
YES
NO
Suspicious content
Authentication module
NOT FAKE
Fragile/semifragile WM detector (authentication)
(Semi)Fragile WM detected ?
YES
NO
Media forensics (ML)
Forged?
YES
NO

Figure 2: Step 1 – Source verification and content authentication.

content. A database of reliable sources is used to determine whether the content comes from a reliable source or not.

(b) **Authentication module.** If the source of the content has been identified as reliable, then an authentication module can be used to try to determine if there are traces of forgery in the content. This step can be carried out either using a fragile or semi-fragile watermark at the source or, if no fragile/semi-fragile watermarking is used, then media forensics tools, mostly based on ML methods, can be used to try to determine if there has been any malicious modification on the content.

If the reliable source of a content is verified and the authentication module determines that it has not been forged, then the content is labeled as legitimate (“not fake”) and a message is sent to the user detailing the analysis made on the content. Otherwise, the content is labeled as “suspicious” and further scrutiny is required in Step 2.

2. **Step 2: Fake news detection and traceability.** Once a content is labeled as suspicious in Step 1, this step tries to determine if the content is really fake news, legitimate, or undetermined. Finally, if a content is labeled as “fake news”, a module tries to identify its source to add it to a black list (database) of unreliable sources.

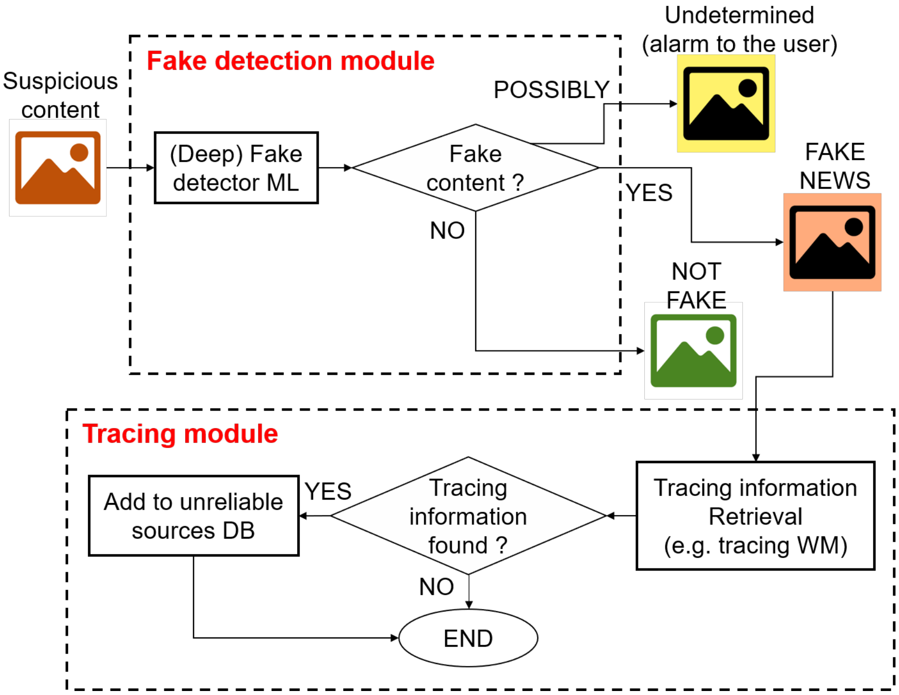
Fake detection module
Suspicious content
(Deep) Fake detector ML
Fake content ?
POSSIBLY
Undetermined (alarm to the user)
YES
FAKE NEWS
NO
NOT FAKE
Tracing module
Add to unreliable sources DB
YES
Tracing information found ?
Tracing information Retrieval (e.g. tracing WM)
NO
END

Figure 3: Step 2 – Fake news detection and traceability.

A flowchart of this phase is shown in Fig.3, where its two modules are outlined:

(a) **Fake detection module.** In this module, a suspicious content is classified using ML tools and three different outputs are possible: legitimate (“not fake”) content, “fake news” or “undetermined”. In the latter case, an alarm will be raised to the user, and he/she will have to decide whether the information can be trusted or not. In addition, if the user wants to, his/her opinion about the reliability of the content will be collected, and it can be used to score the content in case another user sends it for evaluation.

(b) **Tracing module.** If a content is classified as fake news, this module tries to trace the source of the content to try to identify its source and add it to a database of unreliable sources. This step can be accomplished by tracing watermarks (also known as digital fingerprints) or by gathering some other kind of information (provenance data) that may be useful for traceability.

The proposed architecture, as already discussed, makes use of different technologies, including digital watermarking (robust, fragile/semi-fragile, and digital fingerprints), forensic tools, and ML tools trained to discriminate fake news from legitimate contents. This architecture is very flexible, since some

Table 1: Comparison of architecture.

Approach	Methods
Active	Watermark [15] Watermark & Blockchain [46, 33]
Passive	Steganalysis [47, 48] Mesoscopic Network [20] Temporal Domain (LSTM) [49] Frequency Domain [50] etc.
Proposed	Combination of Active & Passive Traceability of source (Digital Fingerprint)

technologies (such as digital watermarks) are not mandatory, but they can significantly help in the detection of fake news if used.

As mentioned in the Introduction (Section 1) there are two distinct types of fake news, namely, those that are created from a legitimate content that is manipulated later on, and the ones that are created as fakes from scratch, without manipulating the original source. The envisaged architecture described in this section aims at facing both types of fake news. The detection process could be the following:

1. Identify the source of the content. If this is feasible, we can have three different outcomes: a) the content is from a known and legitimate source (in that case manipulation is possible to create fake news of the first type); b) the content is from a known source of fake news (in that case the information would be given directly to the user, who could opt to discard the content); and c) the source of the content is undetermined (this would the typical case for fake news of the second type).
2. For known legitimate sources, detect manipulations (fake news of the first type).
3. For unknown sources, use machine learning to classify the content as genuine or fake. Of course, the accuracy of this classification would possibly be low, since in many cases detecting just “lies” would not be easy. A future extension of the project may consider checking whether some piece of news can be found also in a legitimate source. This is what fact-checkers currently do. However, this possibility is out of the scope of the DISSIMILAR project for the time being, due to the limitations of the resources.

At the end of the DISSIMILAR project, we expect to develop a prototype of this architecture and make it open such that other solutions can be integrated, paving the way towards a free-to-use platform to help users in fake news detection. Differing from other solutions, the DISSIMILAR platform will empower media consumers to decide when to use the system to assist him/her in identifying fake news, preventing censorship or centralized control, thus preserving fundamental rights such as the freedom of expression on the Internet.

6.2 Comparative analysis

In the past few years, many researchers focus on the problem of fake content and its defense techniques. The conventional approaches are roughly classified into two types: active and passive. Table 1 shows the comparison of existing methods and the proposed architecture. In active techniques, some information is encoded at the time of multimedia generation, e.g., a watermark is added to the content [15]. The watermark is used to identify if the multimedia content has been manipulated. Furthermore, the manipulated portions in the target can be detected using the extracted watermark. However, digital watermarks are not foolproof [51], and this problem can be countered by incorporating a blockchain [52, 33] to hold a tamper-proof record of watermarks and content features. Alattar et al. [46] has proposed a proof-of-concept fake video news detection and prevention system using watermarking and blockchain technologies.

In passive techniques, some traces of manipulation are analyzed to identify fake contents. DeepFakes frequently produce artifacts that are difficult to identify by humans, but can be recognized by machine and forensic analysis. Inconsistencies, irregularities in the background, and GAN fingerprints are examples of spatial artifacts. Detecting fluctuations in a person's behavior, physiological signals, coherence, and video frame synchronization are all examples of temporal artifacts. The idea of dealing with pixels and exploiting the correlations are one of the straight-forward approaches to clarify the variations between real and fake. To boost the detection efficacy and improve generalization capacity, DNN-based techniques have been investigated in a literature.

The proposed architecture is the combination of active and passive approaches. During the source verification and content authentication, both the existence of watermark and verification of manipulation traces are executed to highly classify the target content whether it is fake or not. In addition to the classification, the proposed architecture involves the tracing module, which enables us to identify its source with the assistance of a digital fingerprinting technique.

7 Project test-bed platform

As described in the preceding paragraphs, the introduced fake news detection methods need a robust and reliable evaluation environment. Moreover, during such an evaluation process, real data gathered directly from the Internet would be required to prove that the developed approaches would be efficient and effective in real-world environments. For this purpose, within the DISSIMILAR project, a special test-bed platform with dedicated custom software will be designed, developed, and deployed. To be more accurate, currently the design of the software platform has already started.

The most essential elements of the designed software are presented in the Fig. 4. The main "workhorse" of the system is the webpage harvester. This subsystem is responsible for visiting selected web pages and downloading multimedia material, for example, digital images, audio, and video files, for further investigation. All downloaded material is stored on the hard drive as well as it is analyzed instantly using the provided plugins (containing, e.g., proposed detection schemes). The developed system allows the easy addition of new analytical plugins that provide various functionalities to the system, for example, detection of embedded watermarks, identification of the possible media modification for fake news creation, or addition of steganographic information for secret sending of commands to the infected machines. During the research, the project partners can easily provide dedicated plugins to test a particular algorithm or method. All results from analytical plugins concerning a given (stored in the hard drive) media file, are placed as metadata in the dedicated database. Due to the fact that this database is essential for the system, it is presented in Fig. 4 as one of the three most crucial parts. The last element of the developed platform is the web-based user interface. It can be used for managing the process of harvesting multimedia data from the Internet. Moreover, it could be used for searching metadata in the system database.

The developed system will be used during various stages of the DISSIMILAR project. During the first phase of the project, it will be used for gathering data from real sources on the Internet. As the project progresses, the developed detection methods, for example, those which would be able to detect the manipulation of multimedia files for fake news, can be added as plugins and evaluated on real-world

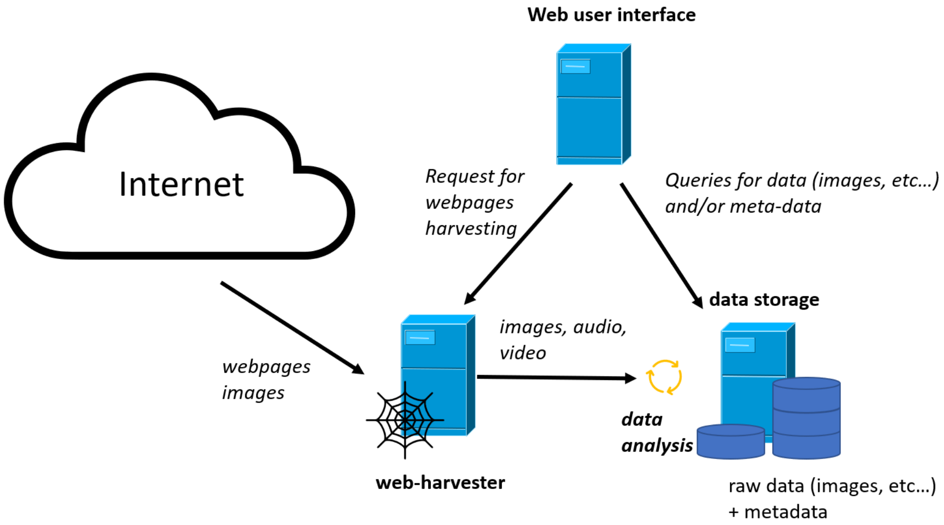
Web user interface
Internet
Request for webpages harvesting
Queries for data (images, etc...) and/or meta-data
webpages images
images, audio, video
data storage
data analysis
web-harvester
raw data (images, etc...) + metadata

Figure 4: Architecture of the test-bed platform.

data.

In the preceding paragraphs, we presented the design of a system that has been used during our experiments. As the conducted research concerns data harvesting from web servers around the globe, some additional experimental aspects should be considered.

The first one is associated with the speed of downloading consequent web pages. From the experimental point of view, the assumed list of web pages should be downloaded as fast as possible. Unfortunately, such behavior could have a negative impact on harvested public webservers. For the webserver administrator, such an activity can be also treated as a kind of Denial of Service (DoS) attack. Moreover, some administrators implement an automatic prevention mechanism that stops providing additional web pages to the address flagged as involved in DoS attacks. On the other hand, the high speed of harvesting could prevent the downloading of some web pages is only one factor. While harvesting vast amounts of web pages, we may visit links to malicious web pages. Often, such hostile domains are blackholed. One possible method for blackholing uses DNS, and in such a solution, blacklisted domains return as response unreachable addresses – for example, 127.0.0.1. The initial experiments confirmed that such behavior exists.

The second most crucial aspect concerns providing information to the web server owners that the harvesting we perform is not a hostile activity, but a research one related to the DISSIMILAR project. Of course, there should be a possibility for web server owners to remove their addresses permanently from the list, which we use for experimental purposes. We investigated various methods for informing web server owners about our research and related activities. The first one utilizes the same IP address used for harvesting as a web server. The hosted web page contains information concerning our research, and a form that can be used by the website owners to permanently remove their addresses from our analyses. The second option utilizes some custom HTTP headers and provides information about the conducted DISSIMILAR-related research and links to web pages with more details. Currently, we have not decided which option is better and would be utilized during the final analysis.

8 Expected impact

The envisioned expected impact of the DISSIMILAR project is threefold:

- For online social media users, this project will provide tools to distinguish between original and altered multimedia content. With these, we expect to enhance the credibility of legitimate news and original content, and minimize the negative effects of fake news. The proposed digital watermarking-based system will allow detecting fake news and their origin, and thus, the awareness of the users of the social media platforms will increase. It will also provide users with tools to manage the potential harmful influence of fake news. With a similar goal, to increase the awareness of the users, the proposed detection tools will also help to detect deep fakes, which are especially difficult to spot and extremely damaging for people appearing in such multimedia contents. The extraction of suitable features from fake contents will be robust against GAN that can fool specific fake detecting methods. Our approach will not be limited to passive signal processing techniques. The combination of active techniques such as data hiding in multimedia content will be a remarkable approach and the multimodal solution will lead a new trend in the detection of fake contents.

- From the content producers' point of view, by tracking authorship in the creation (and, potentially, transformation) of digital content, digital watermarking will help to counter the interaction between untrackable content generation or modification and an authorship factor. Watermarks may prevent a given content producer having to deal with image or trust issues due to fake content attributed to them –and thereby trust issues on digital networks more broadly.

- From a scientific point of view, this project will produce several publications that will improve scientific knowledge in different technical fields (i.e., digital watermarking, ML, and signal processing). Furthermore, publications from the user experience study will provide valuable knowledge on the understanding of the social aspects related to the impact of fake news and their redistribution.

9 Conclusion and Future Work

Fake news distribution via online social platforms has been becoming an increasing problem for the whole societies, and it has been already proved that this may cause real casualties. That is why, in this paper, we present a framework that is designed to detect fake news in the multimedia content. This system is developed within a DISSIMILAR project that has started in June 2021 and is collaboratively executed by an international consortium consisting of three partners from Spain, Poland, and Japan.

The ultimate goal of DISSIMILAR is to create tools which utilize digital watermarking techniques, machine learning, and signal processing, which will enable identifying fake news in social media networks. Additionally, a user experience study is planned to determine, based on the experiences collected from diverse (e.g., in terms of geographical location, gender, age, etc.) users, implications for the design, implementation, integration, and evaluation of the developed tools. Note that this project also aims to increase awareness and attract more research from various academic and industry communities to reduce the influence and spread of fake news among social media platforms. It should be also emphasized that, in DISSIMILAR, we offer an interdisciplinary solution that conveniently merges information hiding methods, machine learning techniques, and multimedia forensics, including analyses of social and cultural challenges related to fake news.

We treat this paper as an initial step toward the above-mentioned goals. In more detail, we focused mostly on extensively characterizing the architecture of the DISSIMILAR framework and the test-bed platform that would be used to evaluate various fake news detection approaches.

As our next step, we plan to produce a DISSIMILAR prototype that will prove that a combination of data hiding, machine learning, and multimedia forensic techniques would create an efficient and effective fake news detection platform that would be more successful than partial solutions on their own. Although such proof-of-concept prototype will be enriched with a first set of digital watermarking and machine learning algorithms with a user-centric design, we also intend to share the created platform to the security community so other researchers and developers are able to contribute with improved and possibly more effective solutions by modifying some of its components.

Our future work envisions creating the main components of the platform, i.e., digital watermarking techniques, detection schemes relying on machine learning, and multimedia forensics. Finally, we plan to implement an evaluation platform that will be able to reliably perform experiments for different detection approaches using real-life data sets of multimedia content.

Acknowledgments

The authors acknowledge the funding obtained from the EIG CONCERT-Japan call to the project Detection of fake newS on SocIal MedIa pLAtfoRms "DISSIMILAR" through grants PCI2020-120689-2 (Spanish Government), JPMJSC20C3 (Japanese Government), and EIG CONCERT-JAPAN/05/2021 (National Centre for Research and Development, Poland). The first author also acknowledges the funding to the project RTI2018-095094-B-C22 "CONSENT" by the Spanish Ministry of Science and Innovation.

References

[1] A. Hunt and G. Matthew. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):21–36, May 2017.

[2] A. Marcus. Cross-cultural user-experience design. In *Proc. of the SIGGRAPH Asia 2011 Courses (SA'11), Hong Kong, China*, pages 1–201. ACM, December 2006.

[3] P. Dourish. Implications for design. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06), Montréal, Québec, Canada*, pages 541—550. ACM, April 2006.

[4] D. Megías, M. Kuribayashi, A. Rosales, and W. Mazurczyk. Dissimilar: Towards fake news detection using information hiding, signal processing and machine learning. In *Proc. of the 16th International Conference on Availability, Reliability and Security (ARES'21), Vienna, Austria*, pages 1–9. ACM, August 2021.

[5] M. Castells. *Communication Power*. Oxford University Press, 2009.

[6] H. Jenkins, S. Ford, and J. Green. *Spreaddable Media, Creating Value and Meaning in a Networked Culture*. New York University Press, 2013.

[7] A. Toffler. *The Third Wave*. William Morrow & Company, 1980.

[8] C. R. Sunstein. *Echo Chambers: Bush V. Gore, Impeachment, and Beyond*. Princeton Digital Books+, 2001.

[9] L. Guo, J. A. Rohde, and H. D. Wu. Who is responsible for Twitter's echo chamber problem? Evidence from 2016 U.S. election networks. *Information Communication and Society*, 23(2):234–251, July 2020.

[10] R. Colmenarejo. La digitalización como paradigma: retos éticos emergentes. In *Diálogos entre ética y ciencias sociales. Teoría e investigación en el campo social.* Editorial Universidad Icesi, 2021.

[11] C. R. Sunstein. *On rumors, How falsehoods spread, why we believe them, and what can be done*. Princeton University Press, 2014.

[12] A. Hunt and G. Matthew. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):21–36, May 2017.

[13] H. Jenkins, M. Ito, and D. Boyd. *Participatory Culture in a Networked Era*. Polity, 2016.

[14] TrustProject. The trust project – news with integrity. `https://thetrustproject.org/` [Online; accessed on March 10, 2022].

[15] D. Megías. Data hiding: New opportunities for security and privacy? In *Proc. of the European Interdisciplinary Cybersecurity Conference (EICC'20), Rennes, France*, pages 1–6. ACM, November 2020.

[16] G. R. Blakley, C. Meadows, and G. B. Purdy. Fingerprinting long forgiving messages. In *Proc. of the Advances in Cryptology (CRYPTO'85), Santa Barbara, CA, USA*, volume 218 of *Lecture Notes in Computer Science*, pages 180–189. Springer Berlin Heidelberg, August 1986.

[17] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, September 1998.

[18] P. Korus and N. Memon. Content authentication for neural imaging pipelines: End-to-end optimization of photo provenance in complex distribution channels. In *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA*, pages 8613–8621. IEEE, June 2019.

[19] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *Proc. of the IEEE International Workshop on Information Forensics and Security (WIFS'18), Hong Kong, China*, pages 1–7. IEEE, December 2018.

[20] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS'18), Hong Kong, China.*, pages 1–7. IEEE, December 2018.

[21] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint*, abs/1812.08685:1–5, December 2018.

[22] R. Tolosana, R. V.-Rodriguez, J. Fierrez, A. Morales, and J. O.-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, December 2020.

[23] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C.-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint*, abs/1910.08854, October 2019.

[24] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2008.

[25] A. Adelsbach and A.-R. Sadeghi. Zero-knowledge watermark detection and proof of ownership. In *Proc. of the International Workshop on Information Hiding (IH'01), Pittsburgh, PA, USA*, volume 2137 of *Lecture Notes in Computer Science*, pages 273–288. Springer Berlin Heidelberg, October 2001.

[26] J. Serra-Ruiz and D. Megías. A novel semi-fragile forensic watermarking scheme for remote sensing images. *International Journal of Remote Sensing*, 32(19):5583–5606, August 2011.

[27] O. Benrhouma, H. Hermassi, A.A. Abd El-Latif, and S. Belghith. Chaotic watermark for blind forgery detection in images. *Multimedia Tools and Applications*, 75(14):8695–8718, July 2016.

[28] J. Serra-Ruiz, A. Qureshi, and D. Megías. Entropy-based semi-fragile watermarking of remote sensing images in the wavelet domain. *Entropy*, 1–21(9):847, August 2019.

[29] D. Megías and A. Qureshi. Collusion-resistant and privacy-preserving P2P multimedia distribution based on recombined fingerprinting. *Expert Systems with Applications*, 71:147–172, April 2017.

[30] M. Kuribayashi and N. Funabiki. Fingerprinting for multimedia content broadcasting system. *Journal of Information Security and Applications*, 41:52–61, August 2018.

[31] M. Kuribayashi and N. Funabiki. Decentralized tracing protocol for fingerprinting system. *APSIPA Transactions on Signal and Information Processing*, 8(1):e2, January 2019.

[32] D. Megías, M. Kuribayashi, and A. Qureshi. Survey on decentralized fingerprinting solutions: Copyright protection through piracy tracing. *Computers*, 9(2):1–26, April 2020.

[33] A. Qureshi, D. Megías, and M. Kuribayashi. Detecting deepfake videos using digital watermarking. In *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, (APSIPAASC'21), Tokyo, Japan*, pages 1786–1793. IEEE, December 2021.

[34] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *Proc. of the 26th European Signal Processing Conference (EUSIPCO'18), Rome, Italy*, pages 2375–2379. EURASIP, February 2018.

[35] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *Proc. of the IEEE Winter Applications of Computer Vision Workshops (WACVW'19), Waikoloa, HI, USA*, pages 83–92, January 2019.

[36] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[37] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes.

In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), California, USA*, pages 38–45. IEEE, June 2019.

[38] P. L. Shrestha, M. Hempel, F. Rezaei, and H. Sharif. A support vector machine-based framework for detection of covert timing channels. *IEEE Transactions on Dependable and Secure Computing*, 13(2):274–283, April 2015.

[39] T. Jung, S. Kim, and K. Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, April 2020.

[40] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'19), Long Beach, CA, USA*, pages 46–52. IEEE, June 2019.

[41] WorldBank. World bank country and lending groups – world bank data help desk, July 2021. `https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups` [Online; accessed on March 10, 2022].

[42] D. Lerch-Hostalot and D. Megías. Unsupervised steganalysis based on artificial training sets. *Engineering Applications of Artificial Intelligence*, 50:45–59, April 2016.

[43] D. Lerch-Hostalot and D. Megías. Detection of classifier inconsistencies in image steganalysis. In *Proc. of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'19), Paris, France*, pages 222–229. ACM, July 2019.

[44] G. Apurva and J. Shomik. Adversarial perturbations fool Deepfake detectors. In *Proc. of the 2020 International Joint Conference on Neural Networks (IJCNN'20), Glasgow, United Kingdom*, pages 1–8. IEEE, July 2020.

[45] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint*, abs/1312.6199:1–10, December 2013.

[46] A. Alattar, R. Sharma, and J. Scriven. A system for mitigating the problem of deepfake news videos using watermarking. *Electronic Imaging*, 2020(4):11701–11710, January 2020.

[47] J. Fridrich and J. Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, May 2012.

[48] D. Cozzolino, G. Poggi, and L. Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In *Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'17), Pennsylvania, Philadelphia, USA*, pages 159–164. ACM, June 2017.

[49] I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, and W. AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Proc. of the 16th European Conference on Computer Vision (ECCV'20), Glasgow, UK*, volume 12356 of *Lecture Notes in Computer Science*, pages 667–684. Springer International Publishing, August 2020.

[50] Q. Yuyang, Y. Guojun, S. Lu, C. Zixuan, and S. Jing. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proc. of the 16th European Conference on Computer Vision (ECCV'20), Part XII, Glasgow, UK*, volume 12357 of *Lecture Notes in Computer Science*, pages 86–103. Springer, Cham, August 2020.

[51] A. Qureshi and D. Megías. Blockchain-based P2P multimedia content distribution using collusion-resistant fingerprinting. In *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPAASC'19), Lanzhou, China*, pages 1606–1615. IEEE, November 2019.

[52] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008. `https://bitcoin.org/bitcoin.pdf` [Online; accessed on March 10, 2022].

Author Biography

David Megías is Full Professor at the Universitat Oberta de Catalunya (UOC), Barcelona, Spain, and also the current Director of the Internet Interdisciplinary Institute (IN3) at UOC. He has authored more than 100 research papers in international conferences and journals. He has participated in different national research projects both as a contributor and as a Principal Investigator. He has also experience in international projects, such as the European Network of Excellence of Cryptology of the 6th Framework Program of the European Commission. His research interests include security, privacy, data hiding, protection of multimedia contents, privacy in decentralized networks, and information security. He is a member of the IEEE.

Minoru Kuribayashi received B.E., M.E., and D.E. degrees from Kobe University, Japan, in 1999, 2001, and 2004. He was a research associate from 2002 to 2007 and an assistant professor from 2007 to 2015 at Kobe University. Since 2015, he has been an associate professor in the Graduate School of Natural Science and Technology, Okayama University. His research interests include multimedia security, digital watermarking, cryptography, and coding theory. He serves as an associate editor for IEEE Signal Processing Letters and Journal of Information Security and Applications, and as a vice chair of the APSIPA Multimedia Security and Forensics Technical Committee. He is a member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He received the Young Professionals Award from the IEEE Kansai Section in 2014 and the Best Paper Award at IWDW 2015 and 2019. He is also a Senior Member of IEEE.

Andrea Rosales holds a degree in Communication and Journalism from *Universidad del Valle*, a Master in Cognitive Systems and Interactive Media and a Ph.D. in Human-Computer Interaction from *Universitat Pomepu Fabra*. Currently, she is Associate Professor of the Communication Studies at *Universitat Oberta de Catalunya* on issues related to communication, technology, and society. Her research focuses on the appropriation of digital technologies by diverse users, and the challenges of the datification of societies and particularly in relation to ageism. Thus, she uses a critical perspective to analyze the impact of digital technologies, and contribute to shaping the intrinsic power relationships in the design of digital technologies.

Krzysztof Cabaj holds M.Sc. (2004), Ph.D. (2009) and D.Sc. (habilitation) (2019) in computer science from Faculty of Electronics and Information Technology, Warsaw University of Technology (WUT). He is currently hired as University Professor at Institute of Computer Science, WUT. Former instructor of Cisco certificated Academy courses: CCNA Routing & Switching, CCNA Security and CCNP at International Telecommunication Union Internet Training Centre (ITU-ITC). His research interests include: network security, honeypots, dynamic malware analysis, data-mining techniques, IoT and Industrial Control Systems security. He is author or co-author of over 70 publications, and supervisor of more than twenty five M.Sc. and B.Sc. degree theses in the field of information security. He took part in over a dozen research projects, among others for EU, ESA, Samsung, US Army and US Air Force. Co-leader of Computer Systems Security Group at Institute of Computer Science.

Wojciech Mazurczyk received the B.Sc., M.Sc., Ph.D. (Hons.), and D.Sc. (habilitation) degrees in telecommunications from the Warsaw University of Technology (WUT), Warsaw, Poland, in 2003, 2004, 2009, and 2014, respectively. He is currently a University Professor with the Institute of Computer Science at WUT and a head of the Computer Systems Security Group. He also works as a Researcher at the Parallelism and VLSI Group at Faculty of Mathematics and Computer Science at FernUniversitaet, Germany. His research interests include bio-inspired cybersecurity and networking, information hiding, and network security. He is involved in the technical program committee of many international conferences, and also serves as a reviewer for major international magazines and journals. From 2016 he is Editor-in-Chief of an open access Journal of Cyber Security and Mobility. Between 2018 and 2021 he served as an Associate Editor of the IEEE Transactions on Information Forensics and Security. He is also a Senior Member of IEEE.