Extending the Scope of Computational Archival Science: A Case Study on Leveraging Archival and Engineering Approaches to Develop a Framework to Detect and Prevent "Fake Video"

Hoda Hamouda, iSchool, University of Brtish Columbia, Vancouver, Canada, hodahamouda@gmail.com

Victoria Lemieux, iSchool, University of British Columbia, Vancouver, Canada, v.lemieux@ubc.ca

Corinne Rogers, Artefactual Systems, Canada, corinne.rogers@gmail.com

Ken Thibodeau, Fordham University, New York, United States, kfthib@gmail.com

Jessica Bushey, North Vancouver Museum and Archives, North Vancouver, Canada, jbusheycontact@gmail.com

James Stewart, Patriot One Technologies, St. John's, Canada, jamess@patriot1tech.com

James Cameron, Patriot One Technologies, St. John's Canada, jamesc@patriot1tech.com

Chen Feng, School of Engineering, University of British Columbia, Kelowna, Canada, chen.feng@ubc.ca

***Abstract*—Thousands of videos are posted online every day. The affordability of video editing tools and social networks has facilitated the creation and spread of videos carrying disinformation, i.e. fake videos. Previous attempts to categorize disinformation have focused on content analysis and ascertaining the intention of creators. To extend these approaches, it is beneficial to incorporate the perspective of other fields that study the trustworthiness of records, such as archival science, to help detect and categorize fake videos. This paper proposes to leverage archival science in combination with computer engineering to devise a new framework for detecting and categorizing fake videos. In doing so, the paper offers a case study of the way in which Computational Archival Science, which blends archival and computational thinking, can be used to contribute to a novel approach towards solving the problem of fake videos.**

Keywords—fake videos, trustworthiness of digital records, authenticity of videos, archival science.

1. INTRODUCTION

Computational Archival Science (CAS) is a blend of both computational thinking and archival thinking. It is defined as "[a]n interdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access" [1]. Although the current definition of CAS focuses mainly on records/archives functions, in this paper we illustrate how CAS can be applied to address wider societal issues, such as detection of fake videos.

To that end, this paper reports on the work in progress relating to a research project which aims at detecting and preventing fake videos by bringing together archival and engineering theory. This research consists of three main phases:

1) generate a classification of fake videos to be able to name their different types;

2) generate a model to detect different types of fake videos; and

3) prototype a solution to protect videos from being "faked" or manipulated.

This paper is reporting on early findings of the first phase.

This research is in line with the stated goal of CAS, which is to "apply the collective knowledge of computer and archival science to understand the ways that new technologies change the generation, use, storage and preservation of records", and to understand "the implications of these changes for archival functions and the societal and organizational use and preservation of authentic digital records" [1]. Thus, this case study aims to contribute to a body of literature and discourse on how a blend of archival thinking and computational thinking may lead to innovative

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

new approaches to societal challenges, such as detecting and preventing disinformation in videos.

2. BACKGROUND

A. Problem statement

Every day five hundred hours of video content are uploaded to YouTube, not to mention Facebook, Twitter and other social media platforms. The accessibility of video editing tools, the low barrier to publishing videos online, and the ease of sharing them has accelerated the spread of videos carrying disinformation, i.e. fake videos. A recent special issue of *Scientific American* was dedicated to exploring "Truth, Lies, and Uncertainty" in different areas, including the information ecosystem. It warned of the threat imposed by social media amplification of "toxic misinformation on an unprecedented scale" [2]. Thus, it is becoming increasingly important to ensure and be able to assess the trustworthiness of what we are seeing and hearing.

In September of 2017, Hurricane Irma hit the Caribbean [3]. Facebook user Hendry Moya Duran posted a thirty second video with a caption "Hurricane Irma" which reached 35 million views, had 829 thousand shares and received over one hundred thousand comments [4]. Up until September 2019, his video was still receiving comments, views and shares. The video was efficient in spreading the word on the effect of this natural disaster, except that it was not what it purported to be. The video actually was originally captured in Uruguay, for a tornado that hit Dolores in April of 2016 [5]. This video demonstrates how an examination of the video content alone can be misleading without a deeper examination of the context of creation –e.g. as embedded in the metadata.. For example, there were no clear indications that the natural disaster featured in the video took place in Uruguay and not the Caribbean. Also, the video's visual and audio content was not fabricated; the video was real footage capturing a natural disaster - just not the disaster it purported to capture. In contrast, an examination of the *context* of the video - for example, as embodied in metadata that a YouTube user used to describe the video, captured in Uruguay, the first time it was posted - allows the viewer to make a determination that it is a fake video. This context in archival science refers to concepts such as the *context of creation* of the video-record (i.e. context surrounding the video creation), its provenance[1] and its subsequent custody. The way that Hendry's video was verified to be a fake involved conducting a reverse search of the video. This led to discovery of the original video published to YouTube on April 2016. A comparison of the contexts surrounding the two videos helped identify the original from the fake video instance. While the impact of the disinformation that Hendry's video disseminates might seem relatively harmless, other fake videos can have greater consequences [7]–[10].

Previous research into detecting and preventing fake videos has tended to focus on addressing the issue by analyzing the content of videos in question. This focus on content is widely used by computer scientists and in engineering [11]–[14]. But what about context, which we define as the circumstances "surrounding materials' creation, receipt, storage, or use, and its relationship to other materials" [15]. Or, as defined by archival diplomatics, "the framework in which a record is created, used, and maintained" [15]. We hypothesize that the relatively sophisticated way in which archival science theorizes about the context, as opposed to content, of recorded human communications, including full motion videos, may enhance existing approaches to detecting fake videos. Thus, in this case study, we lay out a framework for expanding the scope of analysis of full motion videos beyond an examination of the videos' content to an examination of videos' context in order to determine authenticity.

In archival science, identification, analysis and description of the *context* of a record plays an important role in protecting its authenticity, interpreting its meaning, and establishing its evidentiary value [16]. The context of a record must be accurately described, and accessible to those who want to use the record, whether as evidence and/or as a source of information [17]. We see value in applying concepts from archival science, specifically archival diplomatics, which studies the nature of records. Archival science defines a record as "[a] document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference" [18]. Even though many of the full motion videos that may be categorized as "fake" would not qualify as records defined by archival science, our basic proposition is that trustworthy records, that is, those that are accurate, reliable and authentic, represent a "gold standard" against which all other *documents*[2] can be compared. We propose a framework that sets forth a number of tests that map to features that we would expect to find in trustworthy video records as defined in archival diplomatics[3]. When these tests are applied to videos and failed, the results may not, *per se*, indicate that the videos are fake; however, we argue that the test results can be used to provide a confidence rating that indicates to the viewer of a given video that it may not be trustworthy and, indeed, could be fake.

B. Definitions

Given that the terms frequently used in our research have different definitions across different disciplines, we present our definitions here.

[1] Provenance refers to information related to the origins, custody, and ownership of an item, such as a video [6].

[2] A *document* is defined as "[a]n indivisible unit of information constituted by a message affixed to a medium (recorded) in a stable syntactic manner" [19].

[3] Archival diplomatics is defined under 2.B Definitions section.

Video in this paper is defined as a sequence of images called frames, that are visually displayed at a certain speed that is the *frame rate,* that may be accompanied with audio, and that carry metadata. Metadata refers to both *technical metadata* and *descriptive metadata*. The former includes metadata that was automatically captured by the recording device such as location, date, time (i.e. timestamp), and resolution, that are embedded in the original video. The latter refers to data that is typically manually added to the video such as its title, its description, added location, added date, tags (categories that the video belongs to such as sports, or politics).

Disinformation here is defined in accordance with the Handbook for Journalism Education and Training [20] as "information that is false, and the person who is disseminating it knows it is false." The difference between *disinformation* and *misinformation* depends on whether or not the person disseminating the information (e.g. video) knows that it is false information. Misinformation is defined as "information that is false, but the person who is disseminating it believes that it is true" [21].

A fake instance of the video (i.e. a fake video) is defined as a video that disseminates disinformation, and this renders it to be an "untrustworthy video". A video containing disinformation is one that is not what it purports to be, and/or that has been tampered with, and/or presented as original when, in fact it contains falsities that do not exist in the original one. In the case of a fake video, there is more than one video involved, there is the one that is not fake i.e. the *original*, which was manipulated and transformed to produce a second video which is the fake video. We refer in our research to the true video as the *original* video. We define an *original video,* based on archival diplomatics, as the first, complete, and effective video record [22]. We refer to any video produced after the original as an *instance* of the video which may or may not be fake.

The *trustworthiness* of a record (whether in the form of a text or video) is defined here according to *archival diplomatics*, which is a body of knowledge that incorporates and integrates concepts and methods from archival science and diplomatics [23]. *Archival science* studies records regardless of whether or not they are kept in archival institutions [16]. It focuses on identifying, authenticating, preserving long-term preservation of and access to these records regardless of their form (e.g., whether text or video). D*iplomatics* is the "study of the creation, form, and transmission of records, and their relationship to the facts represented in them and to their creator" [24]. The *trustworthiness* of a record (whether in the form of a text or video) in archival diplomatics is based on three concepts: authenticity, reliability, and accuracy, [25]. These concepts are explained under the Theoretical Approach section.

Two additional characteristics of a record, namely *persons*, and *contexts* are relevant to this research. First *persons*; the creation of a record involves several *persons* as illustrated in "Fig. 1". The *writer* is the person who has the capacity and authority to articulate the content of the record [26]. In videos, this would be the person who produced the video. The *author* is the person who has the capacity and authority to issue the record [26], [27]. In videos, this would be the person who published (i.e. issued) the video. The author and writer are often the same. The *creator* of the record is the person in whose *fonds* the (video) record exists [28]. The *fonds* is the entire body of records that the creator accumulates by reasons of her function or activity [29]. The *fonds* in the case of an online video can be, for example, the online channel that contains the entire body of videos a YouTube channel owner created, the *fonds,* in this case, is the YouTube channel, and the record *creator,* in this case, is the owner of this YouTube channel. In the case of a news channel, the news agency is the *creator,* and the *fonds* is the entire body of news videos created by the news agency. The record *creator* does not refer to the person who produced the video, this would be the *writer/ author.* The *originator* is the person who owns the account or electronic address from which the video record has been generated, in videos that is the account from which the video is sent/posted/kept [30].

The *addressee* refers to "individual(s) and/or organisation(s) to which the information in the record was addressed" [31]. In videos, this refers to the audience of the video. Each version of a video (e.g. a video and a near-duplicate of the video) has different *persons* attached to it. In the Uruguay's tornado example, the Uruguayan *author* of the original video, is different from Hendry, the *author* of the fake video.

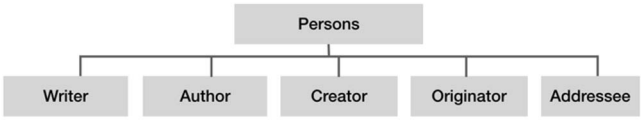

Figure : The creation of a record, in archival diplomatics, involves several *persons*

Second *context;* a record exists in several identifiable contexts, or frameworks, within which the action,[4] that the record participates in takes place [33]. We identify five relevant contexts: juridical-administrative, provenancial, procedural, documentary, and technological [34] as illustrated in "Fig. 2".

[4] In archival science a record, such as a video, must participate in an action, which is "[t]he conscious exercise of will by a person aimed to create, maintain, modify or extinguish situations" [32]. In this research, the action of publishing or posting a video aims at informing the public about an event.

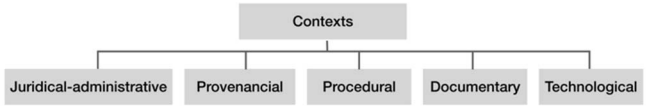

Figure 2: A record, such as a video, in archival diplomatics, exists in five *contexts*

- The *juridical-administrative context* is the legal and organizational system in which the creating body (i.e. *creator)* belongs [35]. In video records this includes, for example, the laws and regulations of the creating body, copyright and licensing rights and restrictions.
- The *provenancial context* refers to the creating body (*creator),* its structure, and functions [36]. This context relates to *provenance,* which is to the relation between records and the organizations or individuals that created, accumulated, or maintained, and used them in the conduct of personal or corporate activity [6]. In videos, this refers to the person or organization who kept it as a record.
- The *procedural context* is the procedure[5] "in the course of which a record is generated [i.e. created]" [33]. This context varies according to the genre of the video. This context can be described in regard to videos published by a news agency, for example, where there are defined steps that the production team follows. It can be challenging to identify a procedural context because sometimes there are no formal or known steps according to which a video is generated.
- The *documentary context* is the archival fonds to which a record (i.e., video) belongs and its internal structure [37], such as the fonds' classification scheme, or video tags–categories that the video belongs to such as sports, or politics. Aspects of the *documentary context* are conveyed through the *extrinsic* and *intrinsic* elements of the *documentary form*[6] [38] of a video record. These *extrinsic elements* include the visual and audio components, as well as the video's frame rate, and audio sample rate [39]. Intrinsic elements of a video include the date, location, description of the event–portrayed in the video–and the name of the writer/author/ originator of the video [39] as illustrated in "Fig. 3". To summarize, the *documentary context* of a video is observed in the creator's *fonds*, its internal structure, and the *extrinsic* and *intrinsic elements* of the *documentary form* of a video. In the case of videos posted to social media platforms, the elements of *documentary form* would be determined by the author/writer rather than the creator.

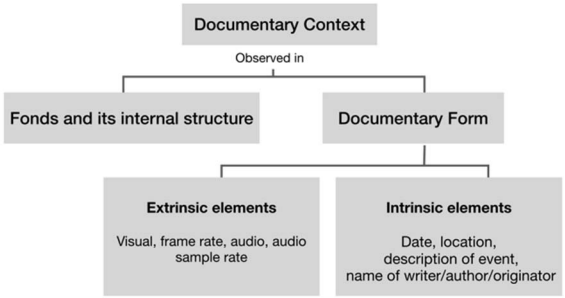

Figure 3: The documentary context of a record is observed in the *fonds'* internal structure and the *documentary form*

- The *technological context* is the "characteristics of the technical components of the electronic system in which the record [such as a video] is created." [40]. This context is related to the software, and hardware systems that process the video, such as the camera, or phone that captured it, and device that was used to upload and publish it. Information about this context can be reflected in the file format of the video, its resolution, and technical metadata that the system or devices that processed video may have captured (e.g. date, geolocation, resolution).

C. *Background Literature on Categorizing and Detecting Fake Information*

There has been much work done on developing a taxonomy for different types of fake information. For example, the synthesis paper by Tandoc, Lim, and Richard [41] on the typology of fake news, Khodabakhsh et al., [11], on audiovisual fake content (i.e. fake videos), the work of Teyssou and Spangenberg on fake video content [42], and the work of Lemieux and Smith on a typology of disinformation in tweets [43]. Tandoc, et al., [41] reviewed how academic articles defined and operationalized the term "fake" in the context of news, i.e. fake news. They examined 34 academic articles published between 2003 and 2017 that referred to the term "fake news". The team concluded that the reviewed academic articles operationalized fake news in six ways, which they referred to as the "typology of fake news", they are: satire, parody, fabrication, photo manipulation, propaganda, and advertising. Their description of the six types is based on two dimensions: (1) the level of facticity, i.e. "the degree to which fake news relies on facts" (however, the authors recognize that an opinion piece is not fake news if the author presents it as an opinion piece) and, (2) the author's intention i.e. "the degree to which the creator of fake news intends to mislead" [41]. Only one out of the six categories was related to visual-based information, i.e., the photo

[5] Procedure is the written or unwritten rules according to which an action (such as publishing a video) is executed, and it consists of formal steps [33].

[6] The *documentary form* refer to the rules that shape the extrinsic and intrinsic elements of a document in order to communicate its content, documentary and administrative context, and its authority [38].

manipulation category. Sadiku et. al., [44] added an additional category to the above mentioned taxonomy, which is "clickbit" (also known as "clickbait"). Another relevant typology is by Khodabakhsh, Busch, & Ramachandra [11]. Their taxonomy is based on the technology used to create fake audiovisual content and focuses on fake videos that involve human subjects. Their study was limited to videos containing talking faces or heads (as opposed to events such as natural disasters). They identify three technological approaches used to generate fake videos: *physical*, *digital*, and *hybrid*. The *Physical* type is when there is a physical human in the video who looks like the original speaker. The *Digital* type is when digital technology such as Computer Graphics Interfaces (CGI) or inter-frame forgeries (i.e. the video or audio are cut and rejoined in a seamless manner), are used to fake a video. The *hybrid* category is when the *author* combines both approaches, such as when a speaker's face is replaced with another using the FaceSwap Application. Teyssou and Spangenberg typology categorized videos based on their content. They identified five types: decontextualized videos, altered decontextualized videos, staged videos, tampered videos, and computer-generated imagery [42]. Lemieux and Smith developed a taxonomy of online disinformation in tweets based on archival theory [43]. The types that the team identified were: misinformation, disinformation, hoax, rumors, hyperbole, bias in terminology, and bias in fact selection.

Two main shortcomings of the above-mentioned approaches motivated our attempt to develop a typology for fake videos: existing typologies do not elaborate on types of fake images and videos. For example, the typology of Tandoc et al., [41] and [44] categorize fake images and videos under one category, "photo manipulation". Moreover, other existing typologies for videos have focused on videos involving talking heads [11], [13], and do not include other genres of videos such as ones related to events (e.g. natural disasters, and protests). Second, we found that one of the dimensions used to categorize information as fake relied upon inferring the intention of the author [41], [44]. This dimension is challenging to apply because it is hard to determine the intention of any actor whose goal is to deceive.

D. Addressing the Gap in Previous Work

To address deficiencies in previous approaches, we have focused our research on genres of videos not covered in previous studies. The types of fake videos we are examining in this paper are ones that have been edited, manipulated, fabricated, or wherein information has been omitted from the visual, audio, and/or metadata of the video, and which singly or in combination result in the video disseminating disinformation. We have also veered away from attempts to guess the intentions of the creator of the video[7]. Our focus is not to classify whether a video is classified as misinformation or disinformation. We have instead focused on testing one or more forms of inconsistency in one or more components of a video. The inconsistency could be:

1) *Visual inconsistency:* when for example an unrelated video, is stitched to the original video
2) *Audio inconsistency:* when for example a part of the audio is cropped or muted
3) *Metadata inconsistency:* when for example the description of the video lacks elements that identify the video as a record or when the existing metadata elements (date, location, format, title, description or video caption) do not correspond to the video's visual and/or audio content.

One category of fake videos that is out of scope of our ongoing research are those that are staged; such as, a video clip in a documentary film on the retail brand "Primark" where, according to investigations, the documentary team asked two girls to pretend to work on the sequins of a t-shirt, so that the documentary could demonstrate that Primark clothing is manufactured by children in India [45]. This type of video, where the videos are classified as fake because parts of the video were acted out or staged, is outside the scope of our research.

3. METHODOLOGY

A. Theoretical Approach

In this section, we present the concept of *trustworthiness* in *archival diplomatics* and its relation to videos. As we have mentioned, a trustworthy digital record, from the point of view of *archival diplomatics,* must meet three criteria, it must be authentic, reliable, and accurate [25] as illustrated in "Fig. 4".

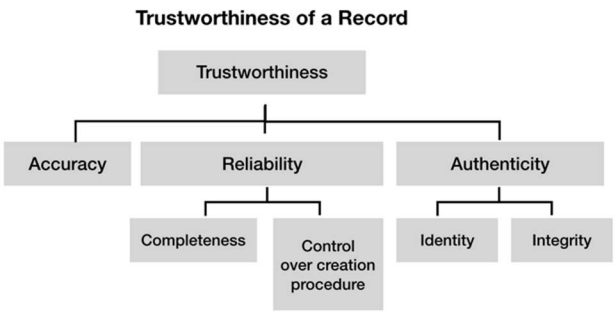

Figure 4 : Trustworthiness of a Record
Source: Adapted from [46]

Accuracy is defined as the degree to which the record "is correct, truthful, and free of error or distortion, whether by omission or commission" [47]. Traditional methods to verify *accuracy* were done through content analysis, but in the digital medium it is harder to presume accuracy using traditional methods [47]. Therefore, in archival science *accuracy* is usually inferred based on the reliability and

[7] It is the human-in-the loop who will infer the intention of the writer/author.

authenticity of a (video) record [43]. On this note, we can state that computer science and engineering approaches to verifying videos fall under verifying the criteria of *accuracy* in *archival science* and *diplomatics* terms.

Reliability is the "trustworthiness of a record as a statement of fact. It exists when a record can stand for the fact it is about". Assessing "the completeness of the record's form and the amount of control exercised on the process of its creation" [46] gives an indication of a record's reliability. This entails according to archival diplomatics that:

- The video record "was created by someone with appropriate authority" [48]. In videos this means that, for example, if a person posts a video of President Obama's speech from a YouTube account that claims to be the White House's official account, this video is not reliable if the person is not affiliated with the White House, and is not someone with official responsibility for posting the video.
- The video record "was made following proper procedures" [48] and was created in the usual and *ordinary course* of conducting business activities [46]. An example would be a video produced by a news agency which has an administratively defined procedure, and therefore is considered to be reliable if the video has been created following these procedures.
- An examination of the system used in the creation of the record and whether it is working as it should be [43]. An example is examining whether the timestamp of the device or system that captured or published the video was accurately set or not.

Authenticity, is the "trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that it is free from tampering or corruption." [49]. It is established by examining the *identity* and *integrity* of the record.

1) *Identity* refers to the attributes that uniquely distinguish one record from other records [49]. It is assessed based on the formal elements of a record and/or its attributes such as the one represented in its metadata [50]. In relation to a video, *identity* includes when and for whom the video was created; the action or matter it participated in and; its (five) relevant contexts [49] which are juridical-administrative, provenancial, procedural, documentary, and technological contexts [15]. Assessing the *identity* is also accomplished through an examination of the *archival bond*. The *archival bond* which is defined as the relationship that links one record to others belonging to and participating in the same activity [51]. In videos this means examining other videos in the *fonds* of the same creator, for example other videos posted by a news channel, or in the case of videos on YouTube, it is examining other videos created by the same user which are published on his/her channel.
2) *Integrity* refers to the (video) record's wholeness and soundness, that it is complete and uncorrupted in all its essential aspects [52] and consequently has the ability to "convey the message it was intended to communicate when generated" [50]. This does not mean that the record must have the same bit structure as it had when created, but means that the message it was meant to convey to achieve its purpose is the same since it was when first created [53].

B. Process of the research

This research began with a three-day workshop that brought together practitioners, academics, and researchers from the disciplines of archival science, digital forensics, computer science, and engineering. The overarching research question of the workshop was: *How to better detect and prevent fake videos?* First, in order to detect fake videos, we found that we needed to work towards building a specification of untrustworthiness in videos, and generalizing a typology of "fakes". In order to derive a typology of fake videos, the research team was presented with twelve case studies wherein videos were considered fake [54]. The subject of the videos varied, some involved political figures and others concerned public officials, natural disasters, or protests.

C. Findings of the workshop

After analysis of these videos, the multidisciplinary team concluded that fake videos can be identified through detection of inconsistencies in one or more components of a video: the visual, audio, or metadata components of a video. These inconsistencies can occur 1) among the components of one video, and/or 2) between the components of two videos, if a near-duplicate video exists.

The team then derived six unique categories of tests that could be used to detect a fake video. The tests relate to three key components of videos: their visual components, their audio components, and their metadata components. The six unique tests are shown in Table 1. Tests 4, 7, and 8 are duplicates of tests 2, 3, and 6 respectively.

TABLE I. Test Matrix for Detecting Fake Videos

Video components of instance 1 of the video	**Video components of instance 1 of a video (and/or if exists a near-duplicate video, i.e. instance 2 of the video)**		
	Visual	***Audio***	***Metadata***
Visual	(1) Tests to check visual against visual VV	(4) Tests to check visual against audio VA (duplicate of 2)	(7) Tests to check visual against metadata VM (duplicate of 3)
Audio	(2) Tests to check audio against visual AV	(5) Tests to check audio against audio AA	(8) Tests to check audio against descriptive

			metadata AM (duplicate of 6)
Metadata	(3) Tests to check descriptive metadata against visual MV	(6) Tests to check descriptive metadata against audio MA	(9) Tests to check metadata against metadata MM

To verify a video, we propose to run tests in two rounds, each consisting of two steps:

- Round 1 is an *internal consistency check* which is a pairwise comparison of the characteristics of each component (visual, audio, metadata) within the same video
- Round 2 is an *external consistency check* which is a pairwise comparison of the characteristics of each component between one instance of a video and another instance of a near-duplicate video if one is available.

Round 1, which we refer to as an *internal consistency check*, consists of two steps.

- In step 1, information about the five contexts of the video, in the form of metadata, is extracted, examined, and compared (such as for example the title, date, location, and author of the video). The metadata is checked to see if there are any inconsistencies within the video itself. As illustrated in "Fig. 5". An example of inconsistency would be if the title of a video states that it was captured in 2019, and the publishing date of the video on YouTube states that it was published in the year 2018.
- In step 2, the video is checked to see if there are any inconsistencies between the metadata (extracted from step 1) and its visual components. An example of inconsistency would be if the title of the video states that it was captured in Cairo, while the visual shows landmarks are from Tunisia. The test also checks for inconsistencies between the metadata and audio, for example if the video was captured in India, while the anchor in the video (i.e. audio component) states that he is in Pakistan.

Both of these steps are illustrated in "Fig. 5".

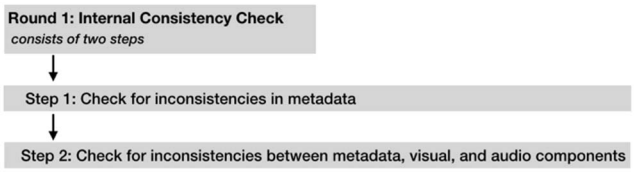

Figure 5: Round 1, the internal consistency check, consists of two steps

Round 2 of the test will take place only if there is a second near identical instance of the video (instance 2) that we refer to as *near-duplicate video* [55]. In this case, each video will first run through the two steps of Round 1, for an *internal consistency check*, then the two videos will run through Round 2 of the test that we refer to as an *external consistency check, (*which is a pairwise comparison of the characteristics of each component between one instance of a video and another instance of a near-duplicate video–if one is available or exists -to identify any alterations that may have occurred from one instance to the next.) It consists of two steps

- In step 1 a comparison between the information available about the five contexts, embodied in the metadata of the two videos will take place. (This refers to test 9 in Table 1, which is a test to check metadata against metadata (MM)).
- In step 2, the test compares the three components of one video (instance 1) and the three components of the other video (instance 2) and searches for inconsistencies between the components (visual, audio, metadata) of the two videos. (This refers to tests 1, 2, 3, 5, and 6 in Table 1.)

Round 2 steps are illustrated in "Fig. 6".

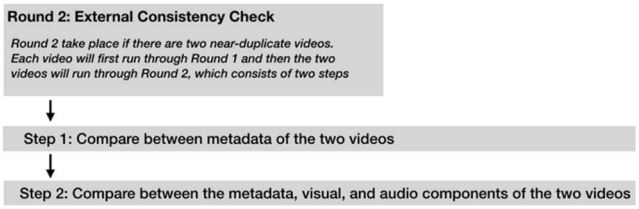

Figure 6: Round 2, the external consistency check, consists of two steps but it takes place only if there are two near-duplicate videos.

The following section presents what each test detects, and the relation between each test and the concept of trustworthiness in archival science and diplomatics (i.e. archival diplomatics[8]). To reiterate, these tests take place within one instance of a video (for internal consistency check) and between two instances of a video, if a *near-duplicate video* is available (for *external consistency check*). The six unique types of tests as shown in Table 1 are:

1- The visual against visual test (VV test number 1). This test is to detect inconsistencies in the visual component(s) of the video(s). Such as
   - Short clips of the video have been cut, or re-attached to another unrelated video.
   - Some clips of the video have been removed.
   - Some parts of the video have been slowed down, or sped up.

[8] Presented earlier under the theoretical framework section.

This test helps assess the *authenticity* (i.e. *identity* and *integrity*) of the video. If visual inconsistencies, such as the ones listed above, are detected this suggests that the *identity* of the record has been tampered with because in this case some *extrinsic elements* of the video (namely the visual images and/or frame-rate of the video) have been altered. If these inconsistencies change the meaning of the message intended to be communicated when the video was generated–then the *integrity*[9] of the video is negatively impacted. An example of a fake video detected by this test is the following: the team working in a Fox News program named the Hannity Show, which used a video to illustrate an interview with republican Michele Bachmann about a Capitol Hill anti "health-care reforms" protest by conservatives that took place on October, 2009 [56]. The team used shots from a second unrelated video along with the original video (i.e. short clips of the video had been cut, and re-attached to another unrelated video.) The second video was from the Taxpayer March on Washington on September 2009, which had a bigger crowd than the October 2009 protest [56]. The video in this case is considered untrustworthy; it is proven inauthentic since it demonstrates inconsistencies in the *identity* of the video record.

2- The visual against the audio test (VA test): This test is to detect inconsistencies between the audio component and the visual component of a video. This could take the form of:
    - Parts of the audio have been cut, or attached to an audio that is not related to the visual. For example, it is not the correct voice over narrative of the reporter.
    - Parts of the audio could have been removed i.e. muted.
    - Parts of the audio could have been slowed down, or sped up.

This test helps assess the *authenticity* (i.e. *identity* and *integrity*) of the video. If inconsistencies between the audio and visual components are detected, such as the ones listed above, this suggests that the *identity* of the record has been tampered with because some *extrinsic elements* of the video (namely the visual images, frame-rate of the video, audio, and/or audio sample rate) have been altered. If these inconsistencies change the meaning of the message intended to be communicated when the video was generated–then the *integrity* of the video is negatively impacted. An example of a fake video detected by this test is a video instance tweeted by Sarah Sanders, from the Infowars site on November 2018. It shows Jim Acosta, CNN's chief White House correspondent, trying to hold the microphone while an intern is trying to take the microphone from him during a press conference with president Trump [7]. Acosta said "Pardon ma'am" to the intern, but the instance of the video edited by Infowars, muted the voice of Acosta [57] i.e. the audio has been removed, therefore the video is untrustworthy, because it is inauthentic. If in the above scenario, muting the voice of Acosta was due to an issue with the audio recording device, this will render the video less *reliable*, rather than inauthentic, because during the creation of the video, the audio system did not function as it should have done.

3- The visual against metadata test (VM). This test is to detect inconsistencies in the metadata of the video in relation to the visual component. This applies when for example:
    - Data descriptions such as the title of the video (or caption), location, date, or the speech of the anchor could be false or inconsistent with the visuals of the video.

This test helps assess the *authenticity*. If inconsistencies between the metadata and visual components are detected, this suggests that the *identity* of the record has been tampered with because in this case some *extrinsic elements* of the video (namely the visual images, frame-rate of the video) and/or some *intrinsic elements* of the video (namely its metadata; title, description, location and/or date) have been altered. If these inconsistencies change the meaning of the message intended to be communicated when the video was generated–then the *integrity* of the video is negatively impacted. An example of a fake video detected by this test is a video instance [58] [55] that was posted as part of a blog by the Washington Post [59]. The title of the article (i.e. metadata) reads "Shocking photos, [from a] video show Egyptian protesters pushing armored police vehicle off bridge". However when watching the video (i.e. visual component), the viewer will see that at "no time could the protesters be seen actually pushing the car off the bridge" [60]. The title and description of the video was inconsistent with the visual, which made the video inauthentic.

4- The audio against metadata test (MA). This test is to detect inconsistencies in the metadata of the video in relation to the audio component. Such as when:
    - Data descriptions such as the title of the video (or caption), location, date, or the speech of the anchor could be false or inconsistent with the audio of the video.

This test helps assess the authenticity of a video because if inconsistencies between the metadata and audio components are detected, this suggests that the *identity* of the record has been tampered with in this case since some extrinsic elements of the video (namely audio) and/or some intrinsic elements of the video (namely its metadata such as title, description, location and/or date) have been altered. If these inconsistencies change the meaning of the message intended

[9] Integrity means that the record is complete and uncorrupted in all its essential aspects [52] and has the ability to "convey the message it was intended to communicate when generated" [50].

to be communicated when the video was generated–then the *integrity* of the video is negatively impacted. An example of a fake video detected by this test is a video instance in which an anchor says (i.e. audio) "I'm standing in front of the pyramids of Giza, in Egypt" but the title of the video (i.e. metadata) reads "A video of the Luxor Hotel and Casino in Las Vegas."

5- Audio against audio test (AA). This test is to detect inconsistencies between the audio components of the video(s).

An example of a fake video detected by this test is a video instance in which an anchor says (i.e. audio) "I'm alone, in a tent in the desert" but there are background acoustics (i.e. audio) of traffic.

6- Metadata against metadata test (MM). This test is to detect inconsistencies between elements of the metadata components of the video(s).

An example of a fake video detected by this test is the video example from the Hannity Show mentioned under the VV test. The title of the video displayed on the screen (i.e. metadata in the form of descriptive title) stated that the video captured the protests of October, 2009, while the date of the unrelated video that was shown (i.e. metadata in the form of date) states that it was captured on September 2009.

If any of these tests for detection of inconsistencies between the different components result in a change in the meaning of the message then the *integrity* of the video is negatively impacted, and thus determined to be inauthentic.

4. CONCLUSION

This case study illustrates the way in which Computational Archival Science can be used beyond records/archival functions. In this case, we apply archival diplomatic theory with the objective of developing an analytic framework to address the societal problem of fake videos. The study illustrates how concepts of archival science, specifically theoretical concepts from archival diplomatics, can be applied to develop six unique tests to detect fake videos. Our research addresses gaps in extent literature on classification of disinformation in three ways: 1) extension to a wider range of video genres, 2) a framework for classifying fake videos that does not rely upon ascription of the motivation of the author of the video or a determination of the author's will in creating the video, and 3) extension of analysis to the context (largely obtained through analysis of video metadata) of the video's creation, i.e., not relying on analysis of the video content alone. To achieve this, our approach proposes a set of tests comprised of pairwise comparisons of the components of a video within the video itself and/or against a near-duplicate video to identify inconsistencies that may signal a lower confidence in the trustworthiness of the video or indicate a potential fake.

Our future work will focus on conducting a human evaluation of our framework to determine whether application of the tests leads human classifiers to more accurately predict whether a video is fake. Based on the results of our evaluation, we will revise our approach and/or our tests to achieve improved results. Once we have undertaken our revisions, we will then design automated techniques to conduct the tests in the context of a human-in-the-loop system that runs the tests as a flag to a human analyst of the possibility that a particular video may be a fake. The goal will be to establish an alert that a human viewer receives indicating that further analysis and investigation may be necessary. The final phase of the research will focus on protecting authentic videos from tampering or detecting tampering that may have occurred. In this way, we aim to combine archival and engineering approaches to achieve a synthesis of both archival and computational thinking that exemplifies CAS.

ACKNOWLEDGMENTS

We would like to thank Dr. Luciana Duranti, Danielle Batista, Dr. Fred Cohen, Darra Hofman, Dr. Zheng Liu, Dr. Sharon Thibodeau, and Dr. Heather O'Brien for feedback on our methodology and the development of this paper.

REFERENCES

[1] R. Marciano *et al.*, "Archival Records and Training in the Age of Big Data," 17-May-2018. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/S0065-28302018000044B010/full/html. [Accessed: 03-Oct-2019]

[2] S. Fletcher, J. Schwartz, and K. Wong, "Truth, Lies & Uncertainty," *Scientific American*, vol. 321, no. 3, Sep-2019 [Online]. Available: http://www.scientificamerican.com/article/truth-lies-uncertainty/. [Accessed: 09-Oct-2019]

[3] BBC, "Hurricane Irma: Caribbean islands left with trail of destruction - BBC News." [Online]. Available: https://www.bbc.com/news/world-latin-america-41218002. [Accessed: 17-Sep-2019]

[4] "Hendry Moya Duran irma - Facebook Search," *https://www.facebook.com/hmoyaduran/videos/1859485954092011/*. [Online]. Available: https://www.facebook.com/search/top/?q=Hendry%20Moya%20Duran%20irma&epa=SEARCH_BOX. [Accessed: 20-Aug-2019]

[5] D. Evon, "FACT CHECK: Is This Hurricane Irma?" [Online]. Available: https://www.snopes.com/fact-check/irma-barbuda-video/. [Accessed: 17-Sep-2019]

[6] "Provenance," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]

[7] J. Finn, "Sanders slammed for retweeting InfoWars host's altered Acosta video," *Mail Online*, 08-Nov-2018. [Online]. Available: https://www.dailymail.co.uk/news/article-6366347/Sarah-Sanders-slammed-retweeting-InfoWars-hosts-DOCTORED-video-Jim-Acostas-karate-chop.html. [Accessed: 20-Aug-2019]

[8] "Firing of Shirley Sherrod," *Wikipedia*. 20-Jun-2019 [Online]. Available: https://en.wikipedia.org/w/index.php?title=Firing_of_Shirley_Sherrod&oldid=902681184. [Accessed: 20-Aug-2019]

[9] S. G. Stolberg, S. Dewan, and B. Stelter, "White House Apologizes to Shirley Sherrod," *The New York Times*, 21-Jul-2010 [Online]. Available:

https://www.nytimes.com/2010/07/22/us/politics/22sherrod.html. [Accessed: 17-Sep-2019]
[10] BBC, "How the Queen clip drama unfolded," 05-Oct-2007 [Online]. Available: http://news.bbc.co.uk/2/hi/entertainment/7030158.stm. [Accessed: 20-Aug-2019]
[11] A. Khodabakhsh, C. Busch, and R. Ramachandra, "A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 372–377.
[12] Z. Wu, N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, Feb. 2015.
[13] R. Ramachandra and C. Busch, "Presentation Attack Detection Methods for Face Recognition Systems," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, Apr. 2017.
[14] K. Sitara and B. M. Mehtre, "Digital video tampering detection: An overview of passive techniques," *Digital Investigation*, vol. 18, pp. 8–22, Sep. 2016.
[15] "Context," *The InterPARES 2 Project Dictionary*. 2012 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf&CFID=17711785&CFTOKEN=11269093. [Accessed: 15-Mar-2019]
[16] T. Thomassen, "Archival Science." Rowman & Littlefield, Lanham, Maryland, pp. 84–86, 2015.
[17] E. Diamond, "The Archivist as Forensic Scientist: Seeing Ourselves in a Different Way," *Archivaria*, vol. 38, no. 0, Jan. 1994 [Online]. Available: https://archivaria.ca/archivar/index.php/archivaria/article/view/12031. [Accessed: 15-Dec-2018]
[18] "Record," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[19] "Document," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[20] C. Ireton and J. Posetti, Eds., *Journalism,'fake news' and disinformation: handbook for journalism education and training*. United Nations Educational, Scientific and Cultural Organization,UNESCO Publishing, 2018.
[21] C. Wardle and H. Derakhshan, "Module 2: Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information.," in *Journalism,'fake news' and disinformation: handbook for journalism education and training*, C. Ireton and J. Posetti, Eds. United Nations Educational, Scientific and Cultural Organization,UNESCO Publishing, 2018.
[22] "original | Society of American Archivists." [Online]. Available: https://www2.archivists.org/glossary/terms/o/original. [Accessed: 04-Oct-2019]
[23] H. MacNeil, "Creating and Maintaining Trustworthy Records in Electronic Systems: Archival Diplomatic Methods," in *Trusting Records: Legal, Historical and Diplomatic Perspectives*, Springer Netherlands, 2000, pp. 86–113 [Online]. Available: https://www.springer.com/gp/book/9780792365990. [Accessed: 01-Jun-2019]
[24] R. Pearce-Moses, Ed., "Diplomatics," *Glossary of Archival And Records Terminology*. Society of American Archivists, Chicago, p. 120, 30-Aug-2005 [Online]. Available: http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf
[25] "Trustworthiness," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[26] "Writer," *The InterPARES Project Terminology Database*, vol. 3, 4 vols. 7, 6, p. 8, 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[27] "Author," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[28] P. Eppard, "CREATOR GUIDELINES Making and Maintaining Digital Materials: Guidelines for Individuals," in *Interpares 2: Experiential, Interactive and Dynamic Records*, L. Duranti and R. Preston, Eds. Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_complete.pdf
[29] R. Pearce-Moses, Ed., "Fonds," *Glossary of Archival And Records Terminology*. Society of American Archivists, Chicago, p. 173, 30-Aug-2005 [Online]. Available: http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf
[30] "Originator," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[31] "Addressee," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[32] "Act," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf&CFID=17711785&CFTOKEN=11269093
[33] L. Duranti, "The Concept of Electronic Record," in *Preservation of the integrity of electronic records*, Dordrecht ; Boston: Kluwer Academic Publishers, 2002, pp. 9–22.
[34] L. Duranti and R. Preston, Eds., *Interpares 2: Experiential, Interactive and Dynamic Records*. Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_complete.pdf
[35] "Juridical-administrative Context," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[36] "Provenancial Context," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[37] "Documentary Context," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[38] "Documentary Form," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[39] L. Duranti, Ed., "Authenticity Task Force. Appendix 1: Template for Analysis.," in *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, Archilab, 2000 [Online]. Available: http://www.interpares.org/book/interpares_book_j_app01.pdf. [Accessed: 07-Oct-2019]
[40] "Technological Context," *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[41] E. C. Tandoc, Z. W. Lim, and L. Richard, "Defining 'Fake News,'" *Digital Journalism*, vol. 6, no. 2, pp. 137–153, Feb. 2018.
[42] D. Teyssou and J. Spangenberg, "Video Verification: Motivation and Requirements," in *Video Verification in the Fake News Era*, V. Mezaris, L. Nixon, S. Papadopoulos, and D. Teyssou, Eds. Cham: Springer International Publishing, 2019, pp. 3–14 [Online]. Available: https://doi.org/10.1007/978-3-030-26752-0_1. [Accessed: 17-Oct-2019]
[43] V. Lemieux and T. D. Smith, "Leveraging Archival Theory to Develop A Taxonomy of Online Disinformation," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4420–4426.

[44] M. Sadiku, T. Eze, and S. Musa, “FAKE NEWS AND MISINFORMATION,” *International Journal of Advances in Scientific Research and Engineering*, vol. 4, pp. 187–190, Jan. 2018.
[45] M. Sweney, “Primark legal chief claims BBC made firm ‘poster boy of child labour,’” *The Guardian*, 16-Jun-2011 [Online]. Available: https://www.theguardian.com/media/2011/jun/16/bbc-primark-child-labour. [Accessed: 06-Aug-2019]
[46] J. T. Tennis and R. Preston, “Terminology Cross-domain. Task Force Report,” in *International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records*, L. Duranti and R. Preston, Eds. Padova, Italy: Associazione Nazionale Archivistica Italiana, 2008, p. 8 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_part_8_terminology_task_force.pdf
[47] R. Pearce-Moses, Ed., “Accuracy,” *Glossary of Archival And Records Terminology*. Society of American Archivists, Chicago, p. 6, 30-Aug-2005 [Online]. Available: http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf
[48] “Reliability,” *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[49] “Authenticity,” *The InterPARES Project Terminology Database*. 2012 [Online]. Available: http://www.interpares.org/ip3/ip3_terminology_db.cfm?letter=r&term=41. [Accessed: 29-Apr-2019]
[50] L. Duranti, “Diplomatics,” *Encyclopedia of Archival Science*. Rowman & Littlefield, pp. 176–180, 2015.
[51] “Archival Bond,” *The InterPARES 2 Project Dictionary*. 2012 [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf&CFID=17711785&CFTOKEN=11269093. [Accessed: 15-Mar-2019]
[52] C. Rogers, “A Literature Review of Authenticity of Records in Digital Systems: From ‘Machine-Readable’ to Records in the Cloud,” *Acervo - Revista do Arquivo Nacional*, vol. 29, pp. 16–44, Jul. 2016.
[53] H. MacNeil, “Providing Grounds for Trust II: The Findings of the Authenticity Task Force of InterPARES,” *Archivaria*, vol. 54, no. 0, pp. 24–58, Jan. 2002.
[54] H. Hamouda, “A Summary of Twelve Case Studies of Fake Videos,” Kelowna, BC, Canada, 22-Aug-2019 [Online]. Available: https://www.slideshare.net/secret/K3zvA51gP46CO0. [Accessed: 12-Nov-2019]
[55] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, “Finding Near-Duplicate Videos in Large-Scale Collections,” in *Video Verification in the Fake News Era*, V. Mezaris, L. Nixon, S. Papadopoulos, and D. Teyssou, Eds. Cham: Springer International Publishing, 2019, pp. 91–126 [Online]. Available: https://doi.org/10.1007/978-3-030-26752-0_4. [Accessed: 17-Oct-2019]
[56] *Sean Hannity Confesses Using Fake Footage: “Jon Stewart Was Right,”* vol. https://www.youtube.com/watch?v=VgOnBUWygsc. [Online]. Available: https://www.youtube.com/watch?v=VgOnBUWygsc. [Accessed: 20-Aug-2019]
[57] Washington Post, *Watch two versions of Acosta video side-by-side - YouTube*. [Online]. Available: https://www.youtube.com/watch?v=aXZ2jRZMLrg. [Accessed: 20-Aug-2019]
[58] *Video of Police Van falling off bridge* فيديو واضح لسقوط المدرعة من كوبري أكتوبر -. [Online]. Available: https://www.youtube.com/watch?v=OeTQJrFMzns. [Accessed: 19-Aug-2019]
[59] M. Fisher, “Shocking photos, video show Egyptian protesters pushing armored police vehicle off bridge,” *Washington Post*, 14-Aug-2013 [Online]. Available: https://www.washingtonpost.com/news/worldviews/wp/2013/08/14/shocking-photos-video-show-egyptian-protesters-pushing-armored-police-vehicle-off-bridge/. [Accessed: 07-Aug-2019]
[60] C. Koettl, “Using UGC in human rights and war crimes investigations,” in *The verification handbook. European Journalism Centre*, 2015 [Online]. Available: https://verificationhandbook.com/book2/chapter7.php