Chapter 1
Verification of web videos through analysis of their online context

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, Ioannis Kompatsiaris

Abstract This chapter discusses the problem of analysing the online "context" of User Generated Videos (UGVs) with the goal of extracting clues that help analysts with the video verification process. As video context, we refer to information surrounding the video i.e. information about the video itself, user comments below the video, information about the video publisher and any dissemination of the video through other video platforms or social media. As a starting point, we present the Fake Video Corpus, a dataset of debunked and verified UGVs that aims at serving as reference for qualitative and quantitative analysis and evaluation. Next, we present a web-based service, called Context Aggregation and Analysis, which supports the collection, filtering and mining of contextual pieces of information that can serve as verification signals. This service aims to assist Internet users in their video verification efforts.

Olga Papadopoulou
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, e-mail: olgapapa@iti.gr

Markos Zampoglou
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, e-mail: markzampoglou@iti.gr

Symeon Papadopoulos
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, e-mail: papadop@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece, e-mail: ikom@iti.gr

1.1 Introduction

User Generated Content (UGC) currently plays a major role in news reporting since publishing information and media content on the Web has become very accessible to Internet users. Recent surveys[1] showed that smartphone users worldwide reached 2.5 billion in 2018. Bystanders, who happen to witness newsworthy events, often act as journalists and share content about the event in their personal social media profiles (e.g. Facebook, Twitter) and to well-known video platforms (e.g. YouTube, Vimeo). This uncontrolled spread of information has exacerbated the challenge of disinformation, also known as "fake news", but has also created a source of important information for news stories, including images and videos, that would otherwise be inaccessible to news organizations. While the term "fake" (typically in association with news) is very popular, it may be misleading in light of the complexity of the problem of online disinformation. In this chapter, we will use the term to refer to inaccurate, decontextualised, misleading or fabricated videos, due to its simplicity and recognition, but the reader should be aware of the nuances of the problem [45].

A top-level categorization of "fake" videos includes: a) tampered videos, which have undergone digital processing, typically with malicious purpose, and b) out of context videos, which are genuine but disseminated with false contextual information. A first step towards understanding the challenge of video verification begins by investigating and analysing existing cases of fake videos from the past. To this end, we collected and annotated the first, to our knowledge, dataset of debunked and verified UGVs, called Fake Video Corpus (FVC). The first version of this dataset consisted of 104 videos, of which 55 were fake and 49 real. In Section 1.3.1, a description of the first release of the FVC [32] is provided together with the methodology that was subsequently followed to extend the dataset to a much larger number of cases and to extend its coverage to multiple platforms. The scale of the latest release of the FVC, called FVC-2018 [31], is large enough ($\sim 5K$) to make it suitable for both qualitative and quantitative evaluations.

Most research on video verification focuses on the development of *video forensics* algorithms [49], which aim to find traces of forgery inside the multimedia content of videos. In addition, there are methods that assess video credibility by analysing the information surrounding the video, e.g. video metadata, the user who posted the video, etc. Following the latter direction, we aim to deal with video verification of UGVs and specifically with the problem of discerning whether a suspect video conveys factual information or tries to spread disinformation –in other words, for the sake of brevity, if the video is "real" or "fake". As a first attempt to assist news professionals with the verification of UGC, we developed the *Context Aggregation and Analysis* (CAA) service, which facilitates the verification of user generated video content posted on social media platforms by collecting and analysing the online context around a video post. Specifically, the information collected directly from the corresponding video platform API is aggregated along with information

[1] From the Pew Research Center (http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ - accessed on April 2019).

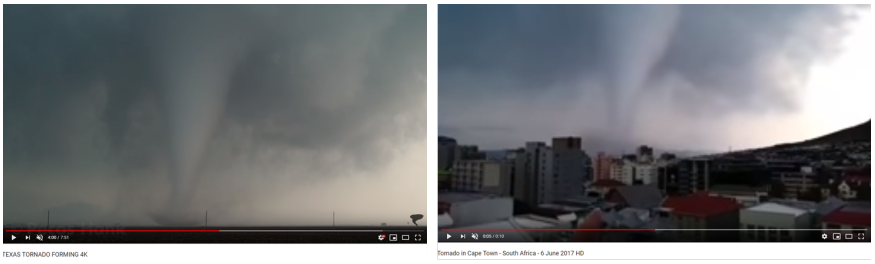


Fig. 1.1: Left: the original footage of a tornado hitting Texas. Right: the doctored video of a tornado hitting Cape Town.

that is extracted or computed by the module to form a first-level verification report. The goal of this tool is to present all the information about the video in one page in order to ease inspection and analysis of the available data by the investigator who is the one in charge of making the final decision regarding the video veracity.

CAA is a tool that aids with the investigation of both tampered and out of context videos. For instance, a tampered fake video was posted on YouTube claiming to show a tornado hitting Cape Town, South Africa. The video circulated on the Web but was finally debunked[2]. In Fig. 1.1, the original footage of a tornado in Texas (left) was used to digitally create the video of a tornado in Cape Town (right). CAA aggregates information about the video and provides indicators that can help investigators decide about the video veracity. One of those indicators derives from reverse image search, which may retrieve the original tornado footage and can provide evidence that there are two videos from different events merged in a single video. Videos associated with misleading contextual information could either be genuine videos from old events that are reposted as breaking news and/or staged videos that are created with intention to mislead, or could be associated with other more nuanced cases of disinformation. For instance, a well known fake video of a Syrian boy running through gunfire and trying to rescue a girl can be debunked with the assistance of the CAA service by leveraging the service facilities for video comment aggregation and filtering (Fig. 1.2). A case of reposting an old video claiming to capture the Brussels airport explosion in 2016, which was actually a video from the Domodedovo airport explosion in 2011, can be debunked with the help of the CAA tool by obtaining pointers to the original footage of Domodedovo (Fig. 1.3).

These are all cases where the CAA tool can assist investigators. Detailed description of the CAA functionalities and usage is provided in Section 1.4. While it is not easy to devise a way to quantify the usefulness of such a tool, some qualitative evaluations of the tool are presented in Section 1.4.2, based on user comments and statistics of the tool usage.

[2] https://www.snopes.com/fact-check/tornadoes-in-cape-town/ - accessed on April 2019.

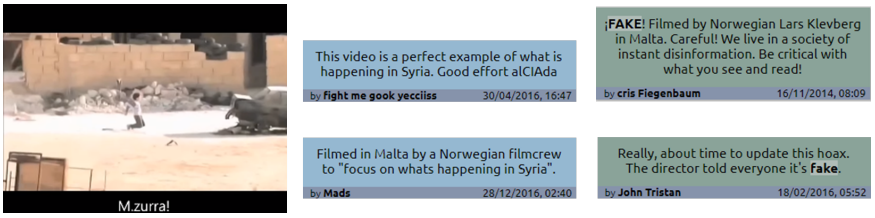


Fig. 1.2: Video depicting a Syrian boy running through gunfire and trying to rescue a girl. Blue: video comments posted below the video. Green: Comments labeled as verification-related.

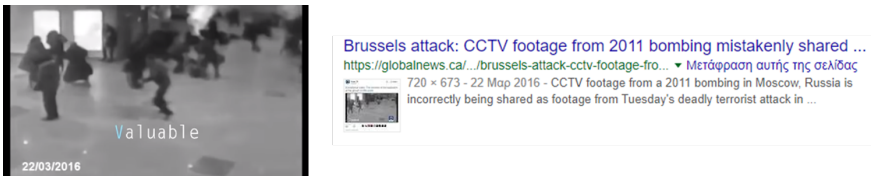


Fig. 1.3: A footage of the Domodedovo airport explosion in 2011 shared as footage of the explosion in Brussels airport in 2016. Reverse image search retrieves references to the original video.

1.2 Related work

The area of multimedia verification consists of various fields of study depending on the type of disinformation under study. A large body of research concerns the detection of traces of forgery and tampering in multimedia items (InVID work in this area is further presented in Chapter 6 of this volume). A family of algorithms, known as "active forensics" attempt to deal with image modifications by embedding watermarks in multimedia content and monitoring their integrity [13, 5]. Another strand of work focuses on searching for telltale self-repetitions [51, 17] or inconsistencies in the image, without making any further assumptions. With respect to image forensics, a recent survey and evaluation of algorithms dealing with these challenges was presented by Zampoglou et al. [49]. The survey explains how these algorithms fail due to limited robustness with respect to image transformation, recompression and rescaling, as it is often the case with social media uploads, where tampering traces quickly disappear as a result of such underlying transformations.

However, there are cases where a multimedia item is used to convey false information not by altering its content, but by presenting it out of its context. In order to confirm or debunk any newsworthy item (text, photo, video), reporters will typically resort to practices such as tracking down the original author that posted the item and ideally contacting them, or looking for inconsistencies between contextual characteristics of the posted item (date, location) and external knowledge about the event. Given the complexity of the problem at hand, there are several multimedia

verification fields of study, tackling various aspects of the problem from different viewpoints.

In the rest of the section, we summarise the strategies, tools and algorithms, which are introduced for dealing with the different types of disinformation. First, an analysis from a journalistic point of view is conducted, concluding that journalists are often cautious of automatic algorithms for content verification. However, they follow guides and tutorials for consulting online tools and decide about an online content veracity (see detailed description in the *Journalistic practices* subsection). Then, a description of semi-automatic and automatic content verification techniques based on *machine learning approaches* is provided. The analysis refers to the extraction of feature/characteristics of the item that is questioned and machine (often deep) learning methods for building verification-oriented classification models. We split this analysis in three categories based on the online item that is questioned; a) rumour analysis, where machine learning practices on hand-crafted features, propagation-based approaches and neural network techniques are examined to detect rumours, b) tweet verification, where features are extracted from the tweet text and the Twitter account that posted it, and c) clickbait content, where similarly characteristics of the post are extracted and machine learning algorithms are used for deciding whether the post is clickbait or not. Finally, the *Verification support* subsection lists fact-checking services and verification tools that exist online and can help with the verification of online content.

1.2.1 Journalistic practices

Automatic verification methods have shown great potential in automatically distinguishing between credible news and disinformation. While journalists generally tend to distrust black-box and fully automatic methods ([40], [50]), preferring to retain a degree of control over the verification process, such tools can provide valuable assistance to them when deciding on the veracity of online content. Journalists are often turning to social networks to extract information and for that reason they need to use verification strategies to verify suspicious social media content and sources [7]. The first attempts to replace ad hoc initiatives with a structured framework consisted of guides and tutorials on what pieces of information are integral for verification, and how to exploit various existing online tools such as Google search[3], reverse image search[4], or Street View[5], for the purposes of verification. One of these reference guides, which is still highly relevant today, is The Verification Handbook by the European Journalism Centre, edited by Craig Silverman[6]. Google News Lab

[3] https://www.google.com/

[4] Google: https://www.google.com/imghp?hl=EN, Yandex: https://yandex.com/images/

[5] https://www.google.com/streetview/

[6] https://firstdraftnews.com/curriculum_resource/the-verification-handbook/

provides its own set of tutorials on how to use Google tools for verification[7] and has also announced its Fact Check feature [28]. The online investigative organization Bellingcat also provides its own guides, including one specific for UGVs[8].

Previous work in the literature try to analyse the behaviour of journalists and the practices that they use to collect and verify UGC from social platforms. For example, Rauchfleisch el al. [36] describe how journalists verify UGC during terrorist incidents, by focusing on the Brussels attacks in March 2016 and the dissemination of UGC through Twitter. With respect to Twitter as a social network for serving and disseminating news content, Heravi et al. [23] analyse the attitudes of journalists in Ireland in order to come up with a set of patterns and practices to serve as global journalistic standards. On the other hand, it is interesting how journalists and social media users approach the various technologies that have been developed for verifying social news content. Looking at the outcomes of a recent work by Brandtzaeg et al. [6], we can see that both professional journalists and citizens are at odds: while some underline the usefulness of the available tools and services for verification, others express significant levels of distrust towards them.

1.2.2 Machine learning approaches

Rumours are pieces of information with truthfulness that is ambiguous or never confirmed. The task of *rumour detection* concerns the accumulation and analysis of a collection of items posted around a claim. According to Zubiaga et al. [53], rumours circulating on social media can be separated in two types: a) long-standing rumours that circulate for long periods of time, and b) newly emerging rumours such as breaking news. Automatic rumour detection methods are categorised by Cao et al. [8] into: a) classification approaches using hand-crafted features, b) propagation-based approaches, and c) approaches based on neural networks. Several studies have been carried out that analyse user behavior, text features and external sources to assess the credibility of a set of tweets comprising a rumour [9, 35, 52, 47]. Moreover, there are approaches that move beyond the extraction of features and focus on modelling the propagation of an event in social networks [48, 29]. In [48], Wu et al. present a novel approach for inferring social media user embeddings from social network structures and utilise an approach based on Long Short Term Memory and Recurrent Neural Networks (LSTM-RNN) for classifying the propagation of news items. With regards to neural networks, RNNs are used by Ma et al. [30] to learn hidden representations of posts, without the need of extracting hand-crafted features. The task of early detection of social media rumours is investigated by Song et al. [38] proposing a model called Credible Early Detection. In contrast to existing methods, which typically need all reposts of a rumour for making the prediction, this work aims to make credible predictions soon after the initial suspicious post.

[7] https://newslab.withgoogle.com/course/verification - accessed on April 2019.

[8] https://www.bellingcat.com/resources/how-tos/2017/06/30/advanced-guide-verifying-video-content/ - accessed on April 2019.

Similarly, previous works on *content verification* rely on extracting characteristics of the text surrounding the multimedia item. In Fig. 1.4, a number of such typical text-based features are presented and categorised in five main groups. A typical case is the work of Boididou et al. [4] where text-based features are used to classify a tweet as "fake" or "real", and show promising results by experimenting with supervised and semi-supervised learning approaches. The approach of Gupta et al. [20] deals with 14 news events from 2011 that propagated through Twitter, and extracts "content-based" (e.g. number of unique characters, pronouns, etc.) and "source-based" (e.g. number of followers, length of username, etc.) features. The approach is evaluated using RankSVM and a relevance feedback approach, showcasing that both groups of features are important for assessing tweet credibility. A comparison of the top performing approaches of the "Verifying Multimedia Use" benchmark task, which took place in MediaEval 2015 [2] and 2016 [3], is presented by Boididou et al. [4] showing promising results in the challenge of automatic classification of multimedia Twitter posts into credible or misleading. The work of Wang et. al [44] presents a large-scale dataset of manually annotated statements from PolitiFact where a hybrid Convolutional Neural Network (CNN) is proposed to integrate metadata with text, showing promising results on the problem of fake news detection. The International Workshop on Semantic Evaluation (SemEval) has introduced tasks dealing with disinformation such as the SemEval-2017 Task 8 'RumourEval: Determining rumour veracity and support for rumours' [15] and the SemEval-2019 Task 4 'Hyperpartisan News Detection' [25]. Both tasks have attracted interest and participation and several machine and deep learning approaches were introduced for dealing with the challenges.

In parallel, there exist a number of techniques designed for detecting *clickbait* content. One such approach is presented by Potthast et al. [34], where 215 features were extracted and evaluated using three classifiers; Logistic Regression, Naive Bayes, and Random Forests. Moreover, the authors presented the first clickbait dataset of tweets. SVMs and Naive Bayes are employed by Chen et al. [11] to tackle clickbait detection using a variety of linguistic features. In [10], Chakraborty et al. present an extensive analysis on linguistic features such as sentence structure, hyperbolic and common phrases, determiners, part of speech tags, etc. The features are evaluated in combination with three different classifiers (SVM, Decision Trees, Random Forests), leading to a 93% accuracy in detecting clickbait. Anand et al. [1] used word embeddings and character level word embeddings as features and an RNN-based scheme as a classifier. RNN and CNNs based on linguistic and network features were used by Volkova et al. [42] to detect satire, hoaxes, clickbait and propaganda.

1.2.3 Verification support

Several means of verification support have been introduced in the recent years to assist journalists and other news professionals to decide on the veracity of news-

related UGC. These can be broadly grouped in two types: a) fact-checking services and b) verification tools. With the spread of unverified information, fact-checking services have gained popularity [19] and based on a Duke Reporters Lab server in 2017 [39], the number of active fact-checking teams was 114. In *automated fact checking* [22], statements are isolated and their veracity is evaluated using reliable databases providing structured knowledge such as FreeBase and DBpedia. Such approaches are generally useful for assessing claims pertaining to historical truths rather than unfolding events. For breaking news, credible fact-checking websites such as FactCheck.org[9], Snopes[10], and StopFake[11] can contribute to the verification. A survey of automatic fact-checking approaches is presented in [41]. Thorne et al. try to unify the definitions presented in related works, which use inconsistent terminology, and identify common concepts by discussing the fact checking in the context of journalism. Then, related works on automated fact checking approaches are collected and organized considering the input, output, and the evidence used in the fact checking process. Social media platforms such as Facebook and Twitter have acknowledged their potential role as means of spreading disinformation, and as a result anti-rumour mechanisms are being designed. Facebook intended to ask users to indicate possible rumours, which would then be sent by the platform to fact-checking organizations such as the AP, FactCheck.org and Snopes.com for verification [24]. Recently, Facebook approached Bloomsbury AI, a London-based startup, with the intention to collaborate and deal together against fake news[12].

With regard to *verification tools*, the image verification tool of Elkasrawi et al. [16] has been proposed to assess the credibility of online news stories by applying a semi-automatic approach using image and text clustering techniques for analysing the image authenticity and consequently the online news story authenticity. The work of Pasquini et al. [33], leveraging the images attached to a news article, tries to identify the visual and semantic similarity between images that appear in articles of the same topic. Fakebox[13] is another tool for verifying news articles by providing information about the title, the content and the domain of the article. The more information is provided, the more accurate the assessment of the article will be. Similarly to Google reverse image search, TinEye[14] supports searches for similar images on the web, which may be useful for journalists when conducting provenance analysis of online video and images.

For *video verification*, Amnesty International's "YouTube Data Viewer"[15] returns the video upload time/date, plus a number of thumbnails (extracted from YouTube) with links to Google reverse image search. Enrique Piracés's Video Vault[16] allows

[9] http://www. factcheck.org

[10] http://snopes.com

[11] http://stopfake.org

[12] https://thenextweb.com/artificial-intelligence/2018/07/02/facebook-just-bought-an-ai-startup-to-help-it-fight-fake-news/ - accessed on April 2019.

[13] https://machinebox.io/docs/fakebox#uses-for-fakebox

[14] http://tineye.com

[15] https://citizenevidence.org/2014/07/01/youtube-dataviewer/ - accessed on April 2019.

[16] https://www.bravenewtech.org/

archiving online videos to save them from being taken down, and provides information in four parts: thumbnails, the metadata of the video as it appeared online, the video footage and audio. It also provides a meeting room where multiple users can share these components and discuss about them in real time. It also allows links for reverse image search on the thumbnails, and a toolbar to slow down playback, speed it up, zoom in on particular areas, rotate the video, and take a snapshot of particular moments.

In a relevant problem, TruthNest[17] and the Tweet Verification Assistant[18] provide solutions for *Tweet verification* using contextual information. In [18], the TwitterTrails web-based tool is introduced in order to help users study the propagation of rumours on Twitter by collecting relevant tweets and important information such as the originator, burst characteristics, propagators and main actors with regard to the rumour. Two more tools have been proposed to help with rumour analysis: Hoaxy [37], which studies the social dynamics of online news sharing, and RumourFlow [14], which integrates visualizations and modelling tools in order to expose rumour content and the activity of the rumour participants.

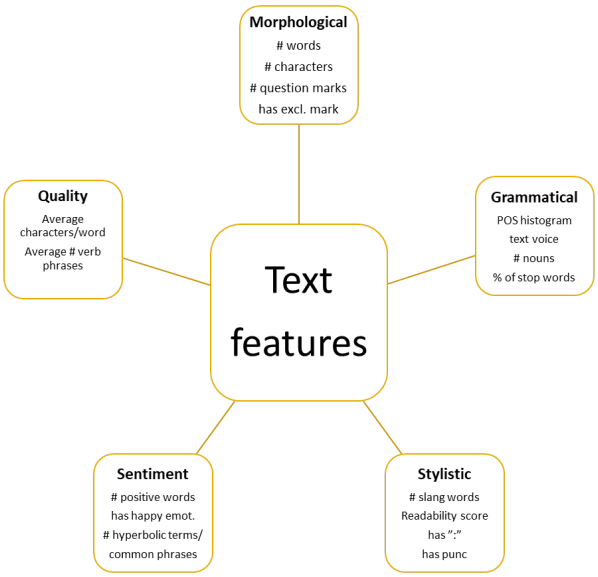


Fig. 1.4: Categorization of text features for characterizing a news item.

[17] http://www.truthnest.com/

[18] http://reveal-mklab.iti.gr/reveal/fake/

1.3 Fake Video Corpus

1.3.1 Dataset collection and overview

The first, to the best of our knowledge, annotated dataset of debunked and verified user-generated videos (UGVs), was created in the context of the InVID project over an extended period of time (2016 - 2018) in cooperation with media experts from the InVID project, and the use of the Context Aggregation and Analysis (CAA) service[19]. The latter is a tool for video verification developed within the InVID project and presented in detail in Section 1.4. The service has drawn attention from the news verification community and has been receiving a large number of video submissions for analysis, which enables the anonymous logging and analysis of videos that were of interest to the verification community.

The first release of the dataset, called Fake Video Corpus (FVC) and introduced by Papadopoulou et al. [32], consisted of 104 videos posted on YouTube, of which 55 were annotated as fake and 49 as real. As the number of cases in this dataset was rather small, the next step was to try and expand the dataset with more fake and real cases. While attempting to expand the dataset, we had to better understand the different types of disinformation. After careful investigation of the literature, we decided to follow the categorisation introduced by Wardle et al. [45]. Below, we provide some examples of videos of the FVC[20] assigned to one of the seven types of mis- and disinformation (Fig. 1.5):

i) **Satire or Parody:** A piece of content obviously intended to amuse viewers that can be misinterpreted as fact, without the intention to cause harm. Example: A video claiming that Pope Francis slapped Donald Trump's hand away, a day after Melania Trump had also slapped his hand away. The latter event was real and was captured in a widely disseminated video. The video with the Pope on the other hand does not show a real interaction between Donald Trump and Pope Francis but was created by the late night television show "Jimmy Kimmel Live".
ii) **Manipulated content:** Content that presents real information but is manipulated in some way to tell a different story. Example: A video with dramatic music and a voice-over decrying migration in Europe, with a shot of the Member of European Parliament Guy Verhofstadt saying "we need migration" in order to frame him as "out of touch" and "dangerous". However, the phrase is cut from an interview with Verhofstadt and taken out of context, removing the nuance of the original statement.
iii) **False connection:** Content that connects two unrelated things, where the video, the caption, or the headline promoting a story does not actually match up with the content. Example: A video claiming that a real dragon was found on a beach. It was watched more than 1.5 million times within three days of its initial posting, and many viewers speculated that the depicted creature was indeed real. The

[19] http://caa.iti.gr

[20] https://mklab.iti.gr/results/fake-video-corpus/

video actually shows a dragon sculpture created for Cuarto Milenio, a Spanish television show.

iv) **False Context:** Genuine content disseminated with false contextual information (taken out of context). Example: A video claiming to depict the moment of an explosion during the attack in Brussels airport in 2016. In truth, it was an older video, shot in 2011 at Domodedovo Airport (Moscow, Russia), and misleadingly shared as footage from the 2016 attack in Brussels.

v) **Fabricated content:** Everything in this type of story is fake and designed with the intention to spread disinformation. It could be either Computer-Generated Imagery (CGI) or staged videos, where actors perform scripted actions under direction, published as UGC. Example: A video supposedly showing a young Syrian boy rescuing a girl amid gunfire. The video was staged, and was in truth filmed by Norwegian Lars Klevberg in Malta.

vi) **Misleading content:** Misleading use of information. Example: A video that purports to show wasted food at a store in Celina, Ohio in fact shows the aftermath of a devastating tornado and the ensuing loss of power. According to the company, due to a tornado the food being disposed of was unsafe to eat after the store lost power for 14 hours.

vii) **Imposter content:** Fake content that purports to come from a real news site or recognised person. Example: A video showing Boko Haram's leader Abubakar Shekau telling his followers to end their violent tactics and to embrace peace. The video turns out to have been fabricated with propagandistic intent.

By analysing the categories listed above and the videos that fall into each one, we can get a glimpse of the breadth and complexity of the problem. On the other hand, there are also real UGVs – that is, videos that convey actual news-related facts with an appropriate context. Collecting such videos from past events is also important in order to have a better perspective of how real and false information gets disseminated and what patterns may distinguish one from the other. Furthermore, since the dataset may also serve as a benchmark for automatic video verification, the inclusion of real (verified) videos is meant to allow evaluations against potential false detections by automatic algorithms (i.e. measure how many real videos will be erroneously classified as fake). From the perspective of creating a representative dataset, this means that the videos need to have been verified first. Videos of which the veracity could not be confirmed with confidence were not included in the dataset. Figure 1.6 presents indicative examples of newsworthy, real videos. The first video shows an extremely rare event of a pod of around 30 Dolphins that were washed ashore and stranded on the beach and then saved by local people at Arraial do Cabo (Brazil). The second video shows a volcano island on the north coast of Papua New Guinea erupting and captured on camera by tourists. The third video, the veracity of which was strongly questioned at the time of its posting, shows a man sitting in his camping chair when a big brown bear walks right up to him and calmly takes a seat.

The need for a dataset more suitable for quantitative evaluation of automatic approaches and the idea to create a large-scale index of fake and real UGVs triggered the extension of the dataset besides more fake and real cases to also more video sources. Between April and July 2017, a second version of the FVC was re-

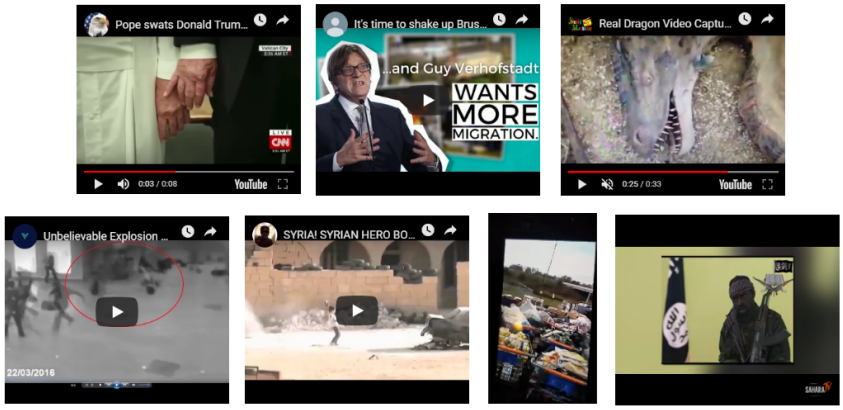


Fig. 1.5: Examples of mis-/dis-information through video. Top: i) The Pope slaps Donald Trump's hand away (Satire/Parody), ii) Guy Verhofstadt is calling for more migration to Europe (Manipulated content), iii) A real dragon found on a beach (False connection). Bottom: iv) Explosion moment in Brussels airport in 2016 (False Context), v) Syrian boy rescuing a girl amid gunfire (Fabricated content), vi) Walmart Throws Away Good Food (Misleading content), vii) Boko Haram leader Abubakar Shekau surrendering on camera (Imposter content).

Fig. 1.6: Examples of real videos. Left: a video of a pod of Dolphins washed ashore and subsequently saved by people; Middle: live footage of a volcano eruption; Right: a bear and a man sitting next to each other.

leased, containing 117 fake videos and 110 real videos. Initially, snopes.com and other debunking sites were consulted in order to collect more debunked fake videos. However, due to the limitations of manually gathering news-related UGVs, a semi-automatic procedure was then followed in order to achieve a larger scale. Between November 2017 and January 2018, all videos submitted to the InVID Context Aggregation and Analysis service were collected, forming a pool of approximately 1600 videos. This set was filtered to remove non-UGC and other irrelevant content, and consecutively, every video within it was annotated as real or fake. In order to annotate the dataset, we used debunking sites to label fake videos, while for real videos we relied on the general consensus from respectable news sources. The resulting FVC-2018, presented in [31], consists of 380 videos (200 "fake", 180

"real"), which were used as basis for retrieving near-duplicate instances of them on three video platforms (i.e. YouTube, Facebook and Twitter). The following steps were executed, resulting into 5,575 near-duplicates of the initial 380 videos:

- For each of the 380 videos, the video title is reformulated in a more general form (called the "event title"). For example, a video with title "CCTV: Manila casino attackers last moments as he enters casino, sets it on fire" was assigned the event title "Manila casino attack".
- The event title is then translated using Google Translate from English into four languages (Russian, Arabic, French, and German). The languages were selected after preliminary tests showed that these were the most frequently appearing in the near-duplicate videos.
- The video title, event title, and the four translations are submitted as search queries to the three target platforms (YouTube, Facebook, Twitter) and all results are aggregated in a common pool.
- Using the near-duplicate retrieval algorithm of Kordopatis-Zilos et al. [27], we filter the pool of videos in order to discard unrelated ones.
- Finally a manual confirmation step is used to remove erroneous results of the automatic method and only retain actual near-duplicates.

The first posted video and all its near-duplicates (temporally ordered by publication time) constitute a *video cascade*. Examples of a real (top) and a fake (bottom) video cascade are presented in Fig. 1.7. During the manual confirmation step, an additional labelling of the near-duplicates of the 200 *fake* videos was applied into the following categories:

- **Fake/Fake:** near-duplicate videos that reproduce the same false claims
- **Fake/Uncertain:** near-duplicate videos that express doubts on the reliability of the fake claim; e.g. the title or the description of the video state that it is fake
- **Fake/Debunk:** near-duplicate videos that attempt to debunk the original claim
- **Fake/Parody:** near-duplicate videos that use the content for fun/entertainment; e.g. by adding funny music effects or slow motion
- **Fake/Real:** near-duplicate videos that contain the earlier, original source from which the fake was inspired.

For the 180 initial *real* videos of the FVC-2018, their near-duplicates were manually assigned to one of the corresponding categories:

- **Real/Real:** near-duplicates videos that reproduce the same factual claims
- **Real/Uncertain:** near-duplicate videos that express doubts on the reliability of the claim
- **Real/Debunk:** near-duplicate videos that attempt to debunk their claims as false
- **Real/Parody:** near-duplicate videos that use the content for fun/entertainment.

In Table 1.1, the number of videos per category is summarized. The category "Private" is a special category assigned only to Facebook videos in cases where the video is posted by a Facebook User or Group, and its context cannot be extracted due to Facebook API limitations. These videos are not further considered in our analysis.

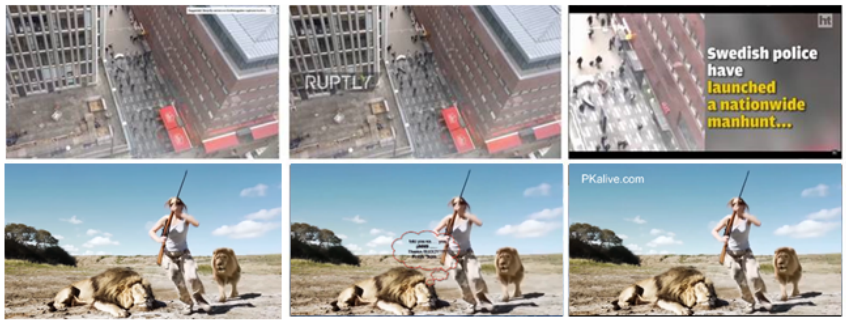


Fig. 1.7: Examples of video cascade. Top: a real video of a truck driving into crowd at a major pedestrian street in central Stockholm, Sweden, on 7 April 2017; Bottom: a fake video of a lion supposedly chasing a trophy hunter to take revenge.

Table 1.1: Categories of real and fake near-duplicate videos collected from YouTube (YT), Facebook (FB) and Twitter (TW). TW shares refer to tweets that share the target YouTube or Facebook videos as a link.)

Fake videos					
	YT	FB	TW	Total	TW shares
Initial	189	11	0	200	-
Fake	1,675	928	113	2,716	44,898
Private	-	467	-	467	-
Uncertain	207	122	10	339	3,897
Debunk	66	19	0	87	170
Parody	43	2	1	46	0
Real	22	51	1	74	0
Total	**2,204**	**1,133**	**125**	**3,462**	**48,965**

Real videos					
	YT	FB	TW	Total	TW shares
Initial	158	22	0	180	-
Real	993	901	16	1,910	28,263
Private	-	350	-	350	-
Uncertain	0	1	0	1	30
Debunk	2	0	0	2	0
Parody	14	6	0	20	0
Total	**1,167**	**930**	**16**	**2,113**	**28,293**

Another step followed to expand the dataset was to submit the URL of the videos of each cascade to Twitter search, and collect all tweets sharing the video as a link. It is a common case, especially for YouTube videos, to be posted on Twitter either as a link or as a link accompanied with text. This step was applied only to the earliest video of each cascade due to the large number of collected tweets (see Table 1.1). The type of Twitter traffic that a video attracts can be a useful indicator of its credibility, but it is a link pointing to a video in the cascade and not another instance of the video. While all types of videos were retained in the FVC-2018 dataset for potential future analysis, the ones considered relevant to the analysis are those which retain the same claims as the initial post, i.e. Fake/Fake and Real/Real. For the rest of this work, all observations and analysis concern exclusively these types of video.

Overall, the scale of the FVC-2018 dataset is comparable to existing datasets for rumour verification. In comparison, the dataset of Gupta et al. [21] contains 16,117 tweets with fake and real images, while the MediaEval 2016 verification corpus contains 15,629 tweets of fake and real images and videos. The data set of Vosoughi

et al. [43] contains 209 rumours with -on average- more than 3,000 tweets each, the collection of which was carried out automatically in order to reach this scale. One important distinction between FVC-2018 and rumour verification datasets is that the FVC-2018 cascades were assembled from disassociated videos using visual similarity, and not from a network of replies or retweets. This is important, since in platforms such as YouTube, such relations between items are not available, making their collection rather challenging.

1.3.2 Dataset Analysis

1.3.2.1 Video and description characteristics

We first analysed the characteristics of the fake and real videos in terms of the videos themselves, their accompanying text and the account that posted them. We compare feature distributions among fake and real videos and present the mean, when normal distribution is followed, or median, otherwise. To further evaluate the statistical significance of the differences between fake and real videos, we compare the mean values using Welch's t-test or the MannWhitneyWilcoxon test and report the associated p-values. Regarding video information, a feature of interest is the video *duration*. The analysis is conducted separately on the first video of each cascade and the overall set of videos in a cascade. We find that, for real videos, the average duration concerning only the first video is 149 seconds and including the near-duplicates the average duration decreases to 124 seconds. On the other hand, for the initial fake videos, the average duration is 92 seconds ($p < 10^{-3}$) and for the cascades 77 seconds ($p < 10^{-3}$). Fake videos tend to be remarkably shorter than real ones.

Concerning the *video poster*, the analysis is conducted on the YouTube channel and the Twitter Users (both native Twitter videos and tweets sharing a video link). Facebook pages are excluded since there is no available information due to Facebook API limitations. First, we examined the age of the channel/account posting the video per video platform, including the near-duplicates. For YouTube real videos, the channel median age is 811 days prior to the day that the video was published, while the corresponding value for fake videos is 425 ($p < 10^{-3}$). The values for Twitter videos are 2,325 and 473 days ($p = 10^{-3}$) respectively. For Twitter shares (tweets containing the link to the initial videos), the difference is minor (1,297 days for real and 1,127 days for fake links) but given the size of the sample it is still statistically significant ($p < 10^{-3}$). Overall, newly created YouTube channels and Twitter users are more likely to post fake videos compared to older accounts. We also find that the YouTube channel subscriber count is 349 users for real videos and 92 ($p < 10^{-3}$) for fake ones. The corresponding value for Twitter accounts is the median follower count of 163,325 users. This particularly high value is due to the fact that only 16 well-established Twitter accounts with many followers were found to have re-uploaded the content as a native twitter video. In contrast, the median number of followers of the Twitter accounts that shared the video as a link is just 333. For fake

videos, the median follower count is 2,855 ($p < 10^{-3}$) for Twitter videos and 297 ($p < 10^{-3}$) for Twitter shares.

Besides user features, linguistic analysis was carried out on the *text* that accompanies the video. Initially the text, specifically the video description for YouTube and Facebook videos and the post text for Twitter, was processed for language detection[21]. For both real and fake videos, the most frequent language is English. However, for fake videos the percentages are lower (see Table 1.2), namely 38% for fake YT videos and 63% for real ones. A high number of posts/descriptions, generally smaller for real videos than fake ones, did not contain enough text for language detection, that is 28% for fake YT videos and 13% for real ones. Other extracted languages which appear at a minor frequency of less than 6% are Russian, Spanish, Arabic, German, Catalan, Japanese and Portuguese, with the exception of Russian fake Twitter videos which are strikingly high (28%).

Building on previous related studies, we studied the following textual features: a) Polarity, b) Subjectivity[22], c) Flesh reading ease ([26]) and d) ColemanLiau index ([12])[23]. Despite the expectation that fake posts should have distinctive linguistic qualities, e.g. stronger sentiment and poorer language [9], no noticeable differences were found between fake and real videos in our dataset (cf. Table 1.2).

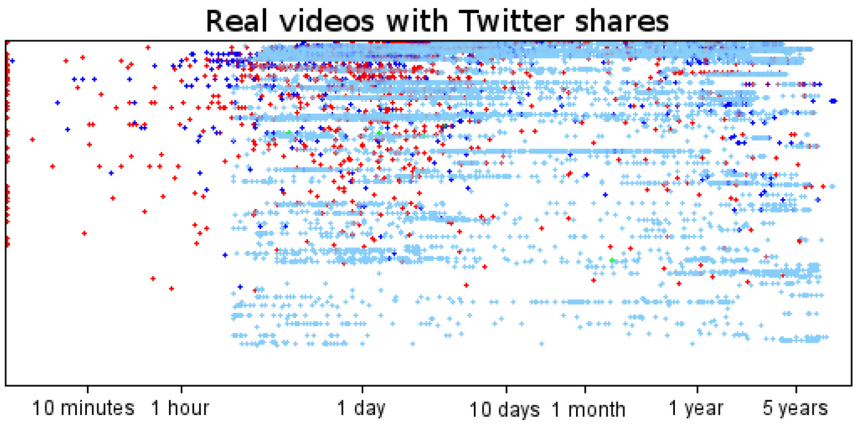
Furthermore, we studied the *temporal distribution* of video cascades. A timeline was created in Fig. 1.8 to show how the near-duplicates of real and fake videos are distributed. Each line corresponds to a video cascade (i.e. the original video and its near-duplicates), while the horizontal axis is the log-scale time between the posting of the initial video and its near-duplicates. Each dot in Fig. 1.8 represents a near-duplicate posted at that time. For clarity, videos are sorted from top to bottom from the most disseminated (more near-duplicate instances) to the least. The time range of near-duplicate spread ranges from a couple of minutes after the initial video was posted up to 10 years; the most important part is that of the difference between fake and real near-duplicates distributions. There are relatively few near-duplicates of real videos posted on YouTube after 10 days from the original post, in contrast to fake videos where near-duplicates are posted at a much higher rate for a much longer interval. This observation also holds for Twitter shares. By calculating the median time difference between the initial video and its near-duplicate, we also confirm this difference. Specifically, for YouTube the median temporal distance is one day for real and 62 ($p < 10^{-3}$) for fake videos, while the values for Facebook videos are 3 and 148 ($p < 10^{-3}$). Regarding Twitter videos, although the values are comparable, one and zero days for real and fake videos respectively, the difference is still significant ($p = 3x10^{-2}$). Finally, for tweets sharing the initial video link, the median distance is 6 and 27 days for real and fake videos, respectively ($p < 10^{-3}$).

A valuable source of information surrounding the videos is the *comments* (or replies in the case of Twitter) that users post below the video. Several comments,

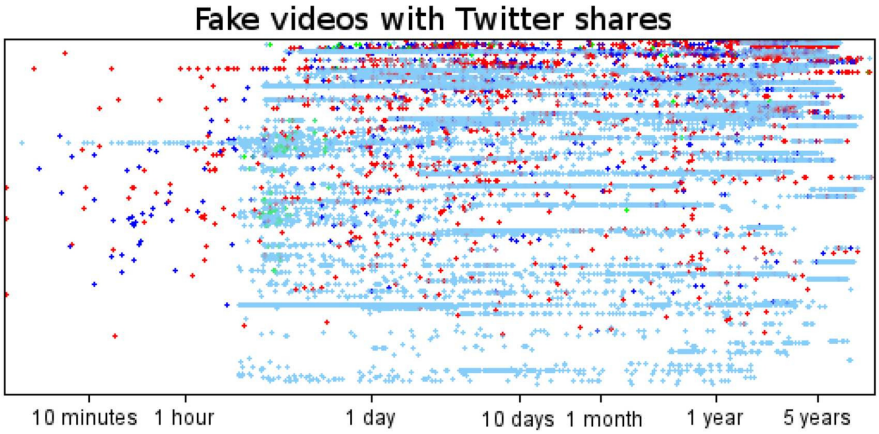
[21] The Python langdetect (https://pypi.org/project/langdetect/) library is used.

[22] The TextBlob Python library (http://textblob.readthedocs.io/en/dev/) was used to calculate Polarity and Subjectivity scores.

[23] The textstat Python library (https://pypi.org/project/textstat/) was used to calculate the Flesh reading ease and ColemanLiau index.



(a) Real videos cascades.



(b) Fake video cascades.

Fig. 1.8: Temporal distribution of video cascades. The near-duplicates are from YouTube (red), Facebook (blue), Twitter (green) and Twitter shares (light blue).

such as the verification-related comments described in Section 1.4, may provide clues that support or refute the video content or claim. Additionally, past work by Papadopoulou et al. [31] offered evidence in favour of the potential of user comments for improving automatic video verification. Overall, 598,016 comments were found on the entire dataset for fake videos, from which 491,639 came from YouTube, 105,816 from Facebook and 561 from Twitter videos. Regarding the real videos, the comments are 433,139 on YouTube videos, 86,326 on Facebook and 215 on Twitter, adding up to a total number of 519,680 comments. Figure 1.9 presents the cumulative average number of comments over time per video for the three video platforms. One may make the following observations: a) a major percentage of comments, es-

	All video platforms (YT, FB, TW)	
	Fake	**Real**
First video duration (seconds)	92	149
All videos duration (seconds)	77	124

	YT		**FB**		**TW**		**Tw shares**	
	Fake	**Real**	**Fake**	**Real**	**Fake**	**Real**	**Fake**	**Real**
Channel/user age (days)	425	811	-	-	2.325	473	1.127	1.297
Subscribers/Followers (#)	92	349	-	-	2.855	163.325	297	333
Video title and description in English (percentage)	38	63	28	41	43	75	52	62
Videos with not enough text (percentage)	28	13	51	48	0	0	4	5
Polarity (float within the range [-1.0, 1.0])	0.091	0.036	0.022	0.056	0.059	0.009	0.078	0.058
Subjectivity (float within the range [0.0, 1.0])	0.390	0.376	0.333	0.307	0.347	0.379	0.452	0.391
Flesh reading ease (float within the range [0, 100])	44.13	37.72	69.27	65.70	35.19	40.89	49.03	48.14
ColemanLiau index (grade level)	15.12	15.01	8.940	11.32	21.64	18.84	17.85	18.22
Time difference (days) between initial video and near-duplicates	62	1	148	3	0	1	27	6
Comments that appear in the first video (percentage)	81	69	22	9	-	-	-	-

Table 1.2: FVC-2018 statistics. The upper table contains the video duration calculated over the first video of the cascades and over all videos of the cascades. The lower table contains statistics per video platform. Dash indicates that there was not enough data for calculating the feature for that video platform (YT: YouTube, FB: Facebook, TW: Twitter).

pecially for YouTube videos, appears in the first video of the cascade with 81% for fake videos against 69% for real, and 22% against 9% for Facebook, respectively; b) the comparison between the number of comments of fake and real videos reveals that the former prevail; c) there is a steep increase in the number of YouTube comments in real videos for a certain period of time (between 12 hours and 10 days after the video is posted), which consecutively tapers off; d) fake videos maintain a steadier rate of accumulation, which, especially after one year from the posting, ends up relatively steeper than for real videos.

The results of the above statistical analysis are summarized in Table 1.2.

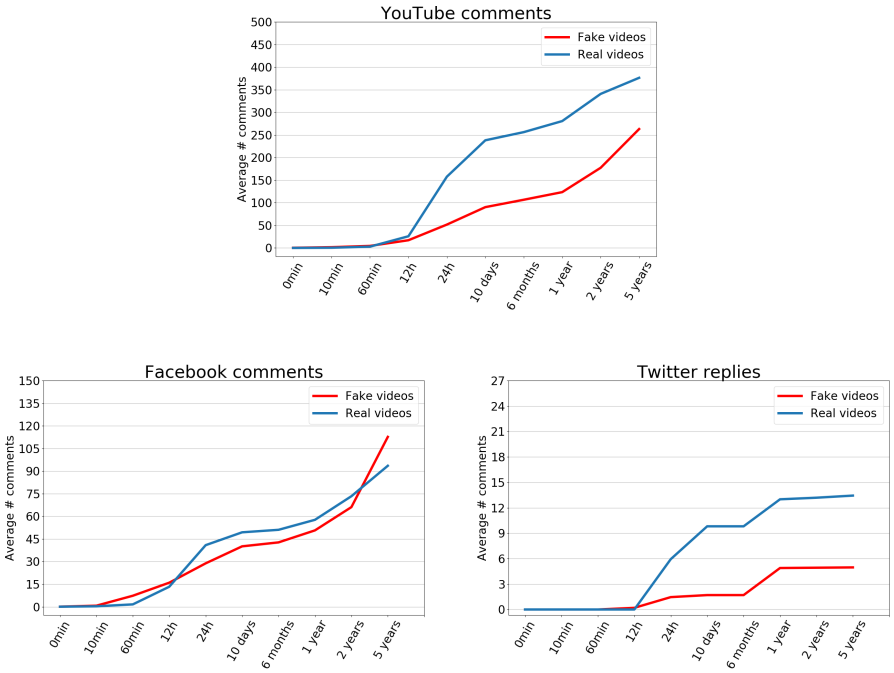


Fig. 1.9: Cumulative average number of comments/replies over time per video for YouTube, Facebook and Twitter.

1.4 Context Aggregation and Analysis

1.4.1 Tool description

The Context Aggregation and Analysis (CAA) tool gathers, filters and summarises several credibility cues to help investigators verify videos shared in online platforms.

In early 2016, the first version of the CAA service was released as part of the InVID project, only supporting the analysis of YouTube videos. The need to extend the tool to more video platforms became apparent following a recent survey from the Pew Research Center[24], which showed that 68% of Americans report that they get at least some of their news on social media, while a majority (57%) say that these news are expected to be largely inaccurate. Given the survey results, YouTube covers a large proportion of the population being informed about news from social networks (23%), but Facebook by far leads with 43%. Twitter follows the other two with 12% and other social media sources, including Instagram, Snapchat, LinkedIn, Reddit, WhatsApp and Tumblr, follow with lower percentages. These observations

[24] http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ - accessed on April 2019.

led us to extend the tool to support Facebook and Twitter videos. In that way, the tool will be useful for the majority of Internet users.

The starting point for using the CAA tool is a URL of a YouTube, Facebook or Twitter video[25]. Then, the tool generates a verification report following the structure of Fig. 1.10. To enrich the verification report, the date and location where the event supposedly happened could be provided as input and the historical weather data of that time and place are included in the report. Overall, the information that the service collects for a submitted video includes:

- Data from source: Information about the video and the channel/user posting the video derived directly from the corresponding video source API.
- Data from Twitter search: Tweets sharing the target video.
- Weather information at the time and place where the event supposedly took place.

The above are provided as input to the different analysis processes of the tool where they are analysed to extract three reports (metadata, twitter timeline, daily and hourly weather) that together make up the overall CAA verification report.

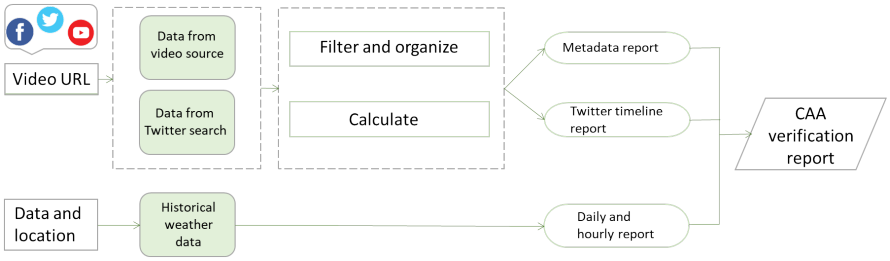


Fig. 1.10: Structure of contextual cues that make up the first-level verification report.

At first, an extensive analysis was carried out over the metadata derived by the APIs of the three video platforms. The amount of data that each video source provides is large and raised the need to carefully filter the information. We concluded to a small, but helpful for verification, list of indicators per video source that are organised in three categories: a) indicators that refer to the video itself (e.g. video title, description); b) indicators providing information about the channel/user that posted the video (e.g. channel description, created time); and c) video comments, where all comments (replies for Twitter videos) posted below the video are aggregated along with the users who posted them and the dates when they were posted. With respect to Facebook, the Graph API is the primary way to get data in and out of Facebook's social graph. There are essentially three types of accounts that may post a video: a) "Facebook User", representing a person on Facebook, b) "Facebook Page", corresponding to a community, organization, business or other non-person entity, and c) "Facebook Group", representing a page where multiple users may post. Facebook

[25] The service supports both Native Twitter videos and tweets with embedded YouTube videos.

User and Group are restricted types and no information can be retrieved for videos posted by these types; some pieces of information are provided by the API only for the case of videos posted by Facebook Pages.

The next step was to map the information across the platforms and create a common reference for all videos. A considerable number of fields could be mapped across all three platforms (e.g. video title and upload time), but there are several indicators that appear in just one or two of them. For example, the "number of times the video has been viewed" is a field provided by the YouTube API, but no such field appears in the Facebook API response; for Twitter, this was mapped to the field "number of times this tweet has been retweeted". For clarity, we created separate metadata reports for videos posted on different video platforms. The video indicators per platform are presented in Fig. 1.11, where on the top are the fields that are common to all video platforms and below are the platform-specific ones. Similarly, Fig. 1.12 illustrates the channel and user indicators for YouTube and Twitter respectively.

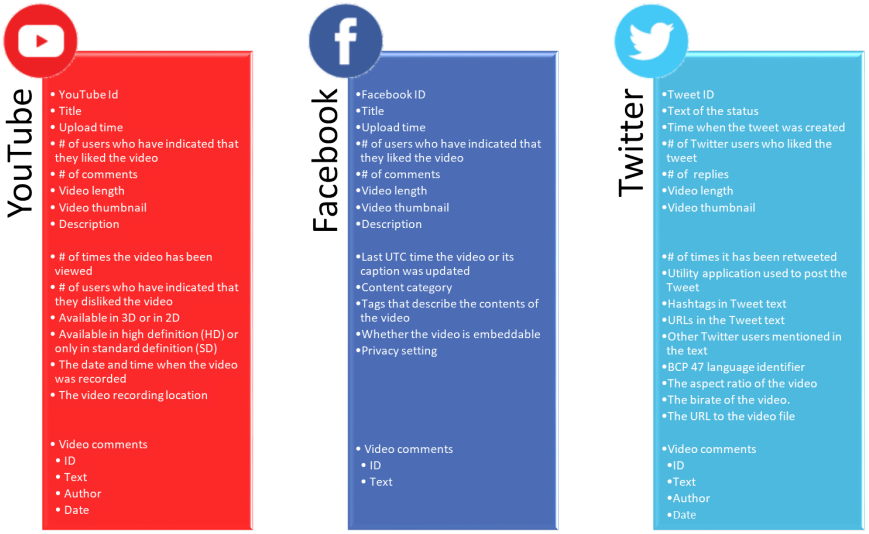


Fig. 1.11: Video verification indicators derived directly by each video platform API. Video indicators refer to information about the video itself.

Part of the report includes a set of credibility indicators that are calculated using the aforementioned metadata fields as presented next.

Verification comments: Users tend to comment on posted videos to express excitement or disappointment about the video content, to share their opinion or a personal experience in relation to what the video shows, and sometimes to doubt or sup-

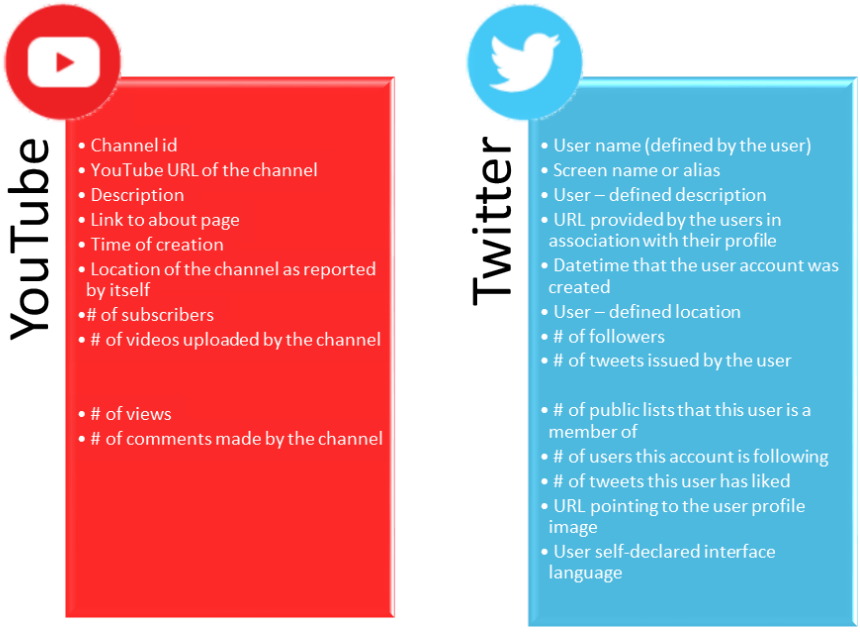


Fig. 1.12: Channel/User verification indicators derived directly by each video platform API. Channel/User indicators refer to the information about the channel (for YouTube) or user (for Twitter) that posted the video. Due to Facebook API limitations, no information about the page posting the video is available.

port the veracity of the video. A list of predefined verification-related keywords[26] is used to filter the comments that may contain useful information for deciding upon video veracity. For example, the fake video entitled "SYRIAN HERO BOY rescue girl in shootout", which claims that a young Syrian boy is rescuing a girl amid gunfire, has a comment which contains the keyword 'fake' and therefore is labeled as a verification-related comment: *"FAKE! Filmed by Norwegian Lars Klevberg in Malta. Careful! We live in a society of instant disinformation. Be critical with what you see and read!"*. The comment explains that the video is staged and shot by a professional film maker. The verification-related list was initially created in English but fake content is disseminated in multiple languages. Figure 1.13 illustrates the fake video of a golden eagle that snatches a kid in a park in Montreal. This video went viral and was spread around the web through multiple video platforms and social media, but also in different languages. To address such cases, we translated the verification-related keywords in six languages (German, Greek, Arabic, French, Spanish and Farsi).

[26] The following list is currently used: 'fake', 'false', 'lie', 'lying', 'liar', 'misleading', 'propaganda', 'wrong', 'incorrect', 'confirm', 'where', 'location'.

Fig. 1.13: The fake video of a golden eagle that snatches a kid in a park in Montreal is disseminated in multiple languages. Actual video titles in English, Spanish and Russian are listed at the right.

Number of Verification comments: The number of verification comments is an important indicator. The higher the number of verification comments, the more likely the video is unreliable or at least worth further investigation.

Locations mentioned: The location where the captured event took place can often provide helpful clues to investigators. Posting a video and claiming that it was captured at a location other than the actual one is a common case of disinformation. In CAA, location extraction from the text metadata of the video are based on Recognyze [46]. Recognyze identifies location-related named entities by searching and aligning them with established knowledge bases such as GeoNames and DBpedia, and refines the results by exploiting structure and context to solve abbreviations and ambiguities, achieving state-of-the-art performance.

Average number of videos per month uploaded by the channel: The number of videos per month is a feature of the channel. The frequency of activity of the channel for real videos is considerably larger than that of fake ones, with the average number of videos per month being 0.018 for fakes and 0.0019 for reals (based on the 380 initial videos of the FVC). Recently created channels posting sensational videos create doubts about the authenticity and reliability of the video. A viral video of a girl being chased by a bear while snowboarding was posted five days after the channel was created. The video gained millions of views before it was debunked[27]. The average number of videos per month uploaded by the channel is calculated by dividing the total number of videos posted by the channel to the number of months that this channel is alive.

Reverse Google/Yandex image search: CAA automatically creates a list of links to easily query Google and Yandex image search engines with the video thumbnails. Apart from the thumbnails that are documented and returned by the YouTube API, there are four additional thumbnails which are automatically constructed by YouTube under the default URL[28]. For YouTube and Twitter the number of thumbnails is fixed, while for Facebook it varies. In cases where the video under consideration is a repost of a previously published video but someone is claiming that it was

[27] https://www.snopes.com/fact-check/snowboarder-girl-chased-by-bear/ - accessed on April 2019.

[28] https://img.youtube.com/vi/youtube_video_id/number.jpg

captured during an unfolding event, reverse image search makes it possible to retrieve the original video and debunk the reposted. Moreover, articles or other videos debunking the video may appear in the results, which could also offer valuable clues.

Twitter search URL: This is a query submitted to Twitter search in order to retrieve tweets that contain a link to the submitted YouTube or Facebook video. With respect to Twitter, the retweets of the submitted tweet are collected.

An additional aggregation step is triggered for each submitted video to collect the ids of the tweets containing a link to that video and use them to generate a Twitter timeline report (an example of which is shown in Fig. 1.14). The tweet IDs can be useful indicators using existing online tools for tweet verification, such as the Tweet Verification Assistant [4]. The Tweet Verification Assistant provides a User Interface[29] that takes a single tweet as input and returns an estimate of the probability that the information contained within the tweet and any associated media (images, videos) is real. It is based exclusively on stylistic features, such as language and punctuation, as well as the profile of the user posting the tweet, and it returns a single value indicating the overall credibility estimate for the tweet as well as the contribution of each individual feature. For Facebook videos, we have experimentally observed that it is in general not such a common case to share them through tweets. Nonetheless, the module searches for tweets sharing the Facebook video and, if they exist, it creates a Twitter timeline report similar to the one created for YouTube videos. With respect to Twitter videos, the retweets of the submitted tweet are similarly used.

Finally, a feature that is calculated only for Twitter videos is included in the verification report. This is a verification label (fake/real), as provided by the Tweet Verification Assistant API.

Another part of the CAA verification report is dedicated to the weather data for a given location and time. We selected the Dark Sky service[30] among the few free online services to obtain weather data. To this end, the CAA module requires as input the time (in the form of a Unix timestamp) and location where the video was supposedly captured. As Dark Sky requires the location in latitude/longitude format, the CAA service converts the location name to lat/lon using the Google Maps service[31] and then submits it to Dark Sky. With respect to the time field, if the exact time (date and time of the day) is given, an hourly report is created for the specific time. Otherwise, if only the date is given, the report refers to the whole day. The Dark Sky service provides various properties but only those relevant to verification are selected by the CAA module.

Figure 1.15 illustrates the daily (left) and hourly (right) weather reports. In both cases, the following features are extracted: a) a short summary of the weather condition, b) the average visibility in km, capped at 10 miles, c) a machine-readable text summary of this data point, suitable for selecting an icon for display. If defined, this property will have one of the values ‘clear-day’, ‘clear-night’, ‘rain’, ‘snow’,

[29] http://reveal-mklab.iti.gr/reveal/fake/

[30] https://darksky.net/

[31] https://developers.google.com/maps/documentation/geolocation/intro - accessed on April 2019.

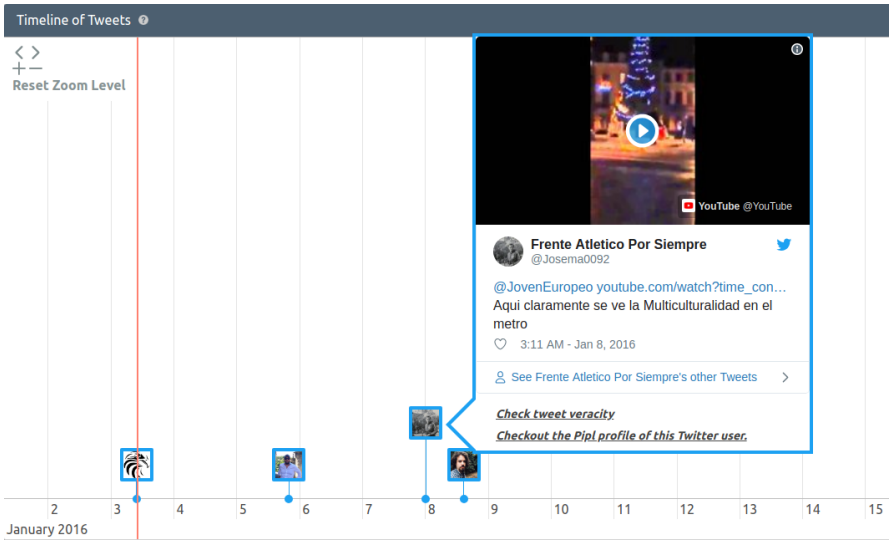


Fig. 1.14: Visualization of twitter timeline report for a YouTube video. A tweet containing the link of the YouTube video was posted couple of minutes after its upload (vertical red line). Three more tweets were posted in the next few days.

'sleet', 'wind', 'fog', 'cloudy', 'partly-cloudy-day', or 'partly-cloudy-night', d) the percentage of sky occluded by clouds, between 0 and 1 (inclusive), e) the wind speed in km per hour and also converted to Beaufort[32], which is a more easy-to-understand measurement, and f) the type of precipitation occurring at the given time. If defined, this property will have one of the values "rain", "snow", or "sleet" (which refers to each of freezing rain, ice pellets, and wintery mix). The different features of these reports refer to the temperature, which for the daily report is a range from maximum to minimum temperature of the whole day, while in the hourly report the exact temperature per hour is provided grouped in three-hour intervals.

Taking into account media experts' and other users' feedback, we extracted two new features. Although the verification comments have proven very useful for the verification process, there are cases where the predefined verification-related keywords are not suitable. Thus, in addition to the predefined keywords and the subset of comments which is automatically extracted by these words, we provide the user with the ability to create a new subset of comments filtered by keywords of their own choice. The user may provide as many keywords as he/she considers to be useful and define a complex boolean query using logical AND/OR operators among the provided keywords.

Finally, considering the FVC-2018 as a rich corpus of fake and real videos, a novel feature of detecting videos that have already been used for (mis-) disinformation is exposed as part of the service. In short, the FVC-2018 dataset includes sev-

[32] https://en.wikipedia.org/wiki/Beaufort_scale

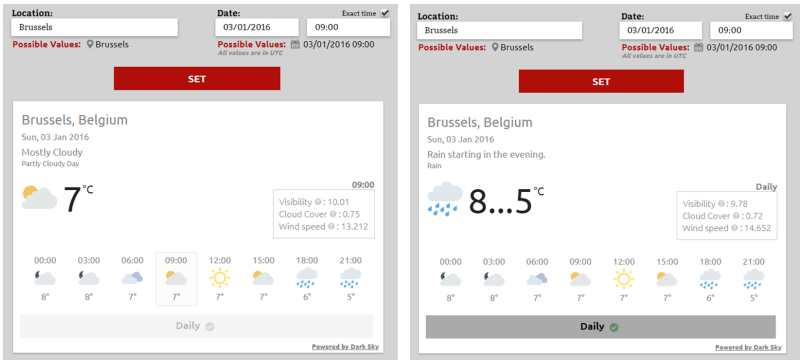


Fig. 1.15: Visual example of the hourly (right) and daily (left) weather report for a certain location. The daytime is split into groups of three hours and the temperature along with an icon indicating the weather condition at that group are presented.

eral cases of debunked videos along with their near-duplicates, accompanied with comments about the false claim, a label indicating the type of the manipulation and optionally a link to an article or a video debunking it. The near-duplicate algorithm of [27] is used to search within the pool of already debunked videos of the FVC-2018. If there is a match, then CAA checks whether the video is unique or has near-duplicates. In the latter case, the URL of the earliest video among all near-duplicate instances is returned. Otherwise, the URL of the matched video is provided. Additionally, as part of the FVC-2018, the accompanying information (i.e. the type of manipulation and the explanation of the false claim) is also included in the report. Some special cases are handled by the CAA service in the following way:

i) The matched video is earlier but has been removed from the video source and is currently not available online. In this case, the report contains the video metadata, specifically the date that it was published and the publisher (channel in the case of YouTube, page in the case of Facebook and user in the case of Twitter video). Moreover, URLs of other near-duplicate instances are provided, if they exist.
ii) The matched video is later than the submitted one. There are cases where more near-duplicates of an event exist but are not part of the FVC-2018 due to the semi-automatic method used to gather the videos. In this particular case, the submitted video is either a near-duplicate which retains the same claim as the matched one or it might be the original video which was later reused to mislead.

The idea of this feature is to protect users to fall again for the same fake videos that was already debunked by reputable sources.

To sum up, the CAA tool does not provide a final decision and does not label the video as fake or real. It creates a verification report that the user should take into account, evaluate its different pieces of information and make the final decision.

1.4.2 Evaluation

The Context Aggregation and Analysis service was evaluated as: a) standalone tool, b) part of the InVID Verification Plugin (Chapter 9 of this volume) and c) part of the InVID Verification Application (Chapter 10 of this volume). In InVID, applications and components are tested and evaluated in various editorial cases and trials. For CAA, tests and evaluations focused on UGVs emerging on YouTube, Facebook and Twitter. Nine Test Cycles were organised and conducted during the InVID project and the CAA service participated in most of them either as a standalone tool or as part of the aforementioned applications. The team of testers consisted of both people with journalistic background and IT specialists for testing the technical interfaces.

In addition to fixing bugs and applying technical refinements, a lot of important feedback was provided by the end users. Some of the most important recommendations include the following:

- The extension to video sources other than YouTube (which was initially the only supported platform) was a strong suggestion. At the second release of the module we covered the most popular and used video platforms in addition to YouTube - Facebook and Twitter.
- Several variations of the input video URLs were noticed. To recognize all the different variations, a preprossessing step was implemented that takes the video URL and extracts the video platform and video id.
- In terms of error handling, the updated version provides more details when an error occurs or when a video cannot be processed.
- The possibility of reprocessing an already submitted video was added. Since new information might become available regarding an online video at any moment (e.g. new comments, likes, etc.), the ability to reprocess the video at any time and collect the new information is essential.
- Performing reverse image search of the video thumbnails was initially supported only using Google image search. After user feedback and investigation of available image search engines, we also included support for Yandex.
- Initially, just the comment/reply text was collected and presented to the user. However, the author of the comment/reply and the date that it was published was proposed as an important clue. This information is available for YouTube and Twitter videos, while the Facebook API does not provide such information.
- With respect to verification comments, requests for additional keywords and multi-language support were taken into account and the verification-related list was extended both in terms of number of keywords and supported languages. Moreover, an additional feature was implemented where the user can define his/her own keywords to perform a comment search.

We use GoAccess[33] and GWSocket[34] for logging and browsing the service statistics. The requests come from direct calls to the CAA API, a UI of CAA developed

[33] https://goaccess.io/

[34] http://gwsocket.io/

for demonstration purposes and the InVID tools that use the CAA API (InVID Verification Plugin and InVID Verification Application). Over a period of 15 months, the service was used by more than 12,000 unique users, from all over the world (United States, France, India, Saudi Arabia and other countries), to assess the veracity of more than 17,000 unique videos.

1.5 Conclusion and Future Work

This chapter presented an analysis of the challenge of verifying online videos, and ways to tackle it using contextual online information. Our starting point was the creation of a large-scale dataset of user-generated videos (200 fake and 180 real videos), along with numerous near-duplicate versions of them that were collected using a semi-automatic process. The value of the dataset to the problem at hand is two-fold: a) it provides a realistic and representative sample of past cases of disinformation based on video content; b) it supports the development of semi-automatic and automatic tools that solve parts of the problem.

Next, we presented the Context Aggregation and Analysis tool, which has been developed within InVID. This collects and analyses the information around an input video and creates a verification report, which aims to assist investigators in their verification efforts. We experimented towards developing a video verification system that could provide the investigator with a direct estimate of whether the video is likely real or fake. Due to the challenge of the problem, we do not have yet an automatic process implemented within the CAA service. The tool is currently applicable to three popular video sharing platforms, YouTube, Facebook and Twitter. However, there are several platforms (e.g. Instagram, WhatsApp, etc.), which are currently widely used or are emerging as sources of eyewitness media. These are not possible to analyze using the CAA service due to limitations or lack of their APIs.

One pertinent issue that we faced during the development of the tool was the challenge of setting up appropriate data collection mechanisms. More often than not, platform APIs did not offer access to information that would be valuable for the task of verification. In addition, during the operation of InVID, Facebook considerably restricted access to their Graph API, as a response to the Cambridge Analytica incident[35]. This considerably reduced the amount of helpful clues that the CAA could collect about the source pages of Facebook videos. Overall, this was another strong case of the well-known Walled Garden issue[36]. The fact that popular Internet platforms such as YouTube and Facebook are in the position to control who has programmatic access to data that is otherwise publicly available makes it very chal-

[35] https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal - accessed on April 2019.

[36] A Walled Garden is a closed ecosystem, in which all the operations are controlled by the ecosystem operator.

lenging to build automated solutions and tools that could help mitigate the problem of disinformation.

Finally, it is worth noting that the problem of disinformation on the Web is much more nuanced compared to a simplistic "fake"-"real" dichotomy. In fact, as became clear by the examples presented in this chapter, several kinds of video-based disinformation abound on the Internet, each with its own particularities.

The presented CAA tool combines existing metadata fields derived directly from the source video platforms along with several calculated indicators, and it aims to generate verification reports, which can be helpful when dealing with most types of disinformation. The tool has been tested by many hundreds of actual end users and its increased use indicates that it is of value to the community of journalists and citizens with interest in verifying multimedia content on the Web. Yet, more research is required along the lines of a) extending the verification report with new indicators and features, and b) making the tool output more easy to digest and interpret by non-trained users.

References

1. Anand, A., Chakraborty, T., Park, N.: We used neural networks to detect clickbaits: You wont believe what happened next! In: European Conference on Information Retrieval, pp. 541–547. Springer (2017)
2. Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y., et al.: Verifying multimedia use at mediaeval 2015. In: MediaEval (2015)
3. Boididou, C., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Middleton, S.E., Petlund, A., Kompatsiaris, Y.: Verifying multimedia use at mediaeval 2016. In: Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016, vol. 1739. CEUR-WS.org (2016)
4. Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y.: Detection and visualization of misleading content on twitter. International Journal of Multimedia Information Retrieval **7**(1), 71–86 (2018)
5. Botta, M., Cavagnino, D., Pomponiu, V.: Fragile watermarking using karhunen–loève transform: the klt-f approach. Soft Computing **19**(7), 1905–1919 (2015)
6. Brandtzaeg, P.B., Følstad, A., Chaparro Domínguez, M.Á.: How journalists and social media users perceive online fact-checking and verification services. Journalism Practice **12**(9), 1109–1129 (2018)
7. Brandtzaeg, P.B., Lüders, M., Spangenberg, J., Rath-Wiggins, L., Følstad, A.: Emerging journalistic verification practices concerning social media. Journalism Practice **10**(3), 323–342 (2016)
8. Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., Li, J.: Automatic rumor detection on microblogs: A survey. arXiv preprint arXiv:1807.03505 (2018)
9. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 675–684. ACM (2011)
10. Chakraborty, A., Paranjape, B., Kakarla, S., Ganguly, N.: Stop clickbait: Detecting and preventing clickbaits in online news media. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 9–16. IEEE Press (2016)
11. Chen, Y., Conroy, N.J., Rubin, V.L.: Misleading online content: Recognizing clickbait as false news. In: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, pp. 15–19. ACM (2015)
12. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2), 283 (1975)
13. Dadkhah, S., Manaf, A.A., Hori, Y., Hassanien, A.E., Sadeghi, S.: An effective svd-based image tampering detection and self-recovery using active watermarking. Signal Processing: Image Communication **29**(10), 1197–1210 (2014)
14. Dang, A., Moh'd, A., Milios, E., Minghim, R.: What is in a rumour: Combined visual analysis of rumour flow and user activity. In: Proceedings of the 33rd Computer Graphics International, pp. 17–20. ACM (2016)
15. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G.W.S., Zubiaga, A.: Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:1704.05972 (2017)
16. Elkasrawi, S., Dengel, A., Abdelsamad, A., Bukhari, S.S.: What you see is what you get? automatic image verification for online news content. In: Document Analysis Systems (DAS), 2016 12th IAPR Workshop on, pp. 114–119. IEEE (2016)
17. Ferreira, A., Felipussi, S.C., Alfaro, C., Fonseca, P., Vargas-Munoz, J.E., dos Santos, J.A., Rocha, A.: Behavior knowledge space-based fusion for copy–move forgery detection. IEEE Transactions On Image Processing **25**(10), 4729–4742 (2016)
18. Finn, S., Metaxas, P.T., Mustafaraj, E.: Investigating rumor propagation with twittertrails. arXiv preprint arXiv:1411.3550 (2014)
19. Graves, L., Nyhan, B., Reifler, J.: Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. Journal of Communication **66**(1), 102–138 (2016)

20. Gupta, A., Kumaraguru, P.: Credibility ranking of tweets during high impact events. In: Proceedings of the 1st workshop on privacy and security in online social media, p. 2. ACM (2012)
21. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web, pp. 729–736. ACM (2013)
22. Hassan, N., Adair, B., Hamilton, J.T., Li, C., Tremayne, M., Yang, J., Yu, C.: The quest to automate fact-checking. world (2015)
23. Heravi, B.R., Harrower, N.: Twitter journalism in ireland: Sourcing and trust in the age of social media. Information, Communication & Society **19**(9), 1194–1213 (2016)
24. Jamieson, A., Solon, O.: Facebook to begin flagging fake news in response to mounting criticism. the Guardian (2016)
25. Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., Potthast, M.: SemEval-2019 Task 4: Hyperpartisan News Detection. In: Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019). Association for Computational Linguistics (2019)
26. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
27. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-duplicate video retrieval with deep metric learning. In: Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on, pp. 347–356. IEEE (2017)
28. Kosslyn, J., Yu, C.: Fact check now available in google search and news around the world (2017)
29. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Aspects of rumor spreading on a microblog network. In: International Conference on Social Informatics, pp. 299–308. Springer (2013)
30. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: IJCAI, pp. 3818–3824 (2016)
31. Papadopoulou, O., Zampoglou, M., Papadopoulos, S., Kompatsiaris, I.: A corpus of debunked and verified user-generated videos. Online Information Review (2018)
32. Papadopoulou, O., Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Web video verification using contextual cues. In: Proceedings of the 2nd International Workshop on Multimedia Forensics and Security, pp. 6–10. ACM (2017)
33. Pasquini, C., Brunetta, C., Vinci, A.F., Conotter, V., Boato, G.: Towards the verification of image integrity in online news. In: Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on, pp. 1–6. IEEE (2015)
34. Potthast, M., Köpsel, S., Stein, B., Hagen, M.: Clickbait detection. In: European Conference on Information Retrieval, pp. 810–817. Springer (2016)
35. Qazvinian, V., Rosengren, E., Radev, D., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 1589–1599. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
36. Rauchfleisch, A., Artho, X., Metag, J., Post, S., Schäfer, M.S.: How journalists verify user-generated content during terrorist crises. analyzing twitter communication during the brussels attacks. Social Media+ Society **3**(3), 2056305117717,888 (2017)
37. Shao, C., Ciampaglia, G.L., Flammini, A., Menczer, F.: Hoaxy: A platform for tracking online misinformation. In: Proceedings of the 25th international conference companion on world wide web, pp. 745–750. International World Wide Web Conferences Steering Committee (2016)
38. Song, C., Tu, C., Yang, C., Liu, Z., Sun, M.: Ced: Credible early detection of social media rumors. arXiv preprint arXiv:1811.04175 (2018)
39. Stencel, M.: International fact checking gains ground, duke census finds. duke reporters' lab, duke university, durham, nc, feb. 28, 2017
40. Teyssou, D., Leung, J.M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O., Mezaris, V.: The invid plug-in: web video verification on the browser. In:

Proceedings of the First International Workshop on Multimedia Verification, pp. 23–30. ACM (2017)
41. Thorne, J., Vlachos, A.: Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:1806.07687 (2018)
42. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, pp. 647–653 (2017)
43. Vosoughi, S., Mohsenvand, M., Roy, D.: Rumor gauge: Predicting the veracity of rumors on twitter. ACM Transactions on Knowledge Discovery from Data (TKDD) **11**(4), 50 (2017)
44. Wang, W.Y.: " liar, liar pants on fire": A new benchmark dataset for fake news detection (2017)
45. Wardle, C., Derakhshan, H.: Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe report, DGI (2017) **9** (2017)
46. Weichselbraun, A., Kuntschik, P., Brasoveanu, A.M.P.: Mining and leveraging background knowledge for improving named entity linking. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, WIMS 2018, Novi Sad, Serbia, June 25-27, 2018, pp. 27:1–27:11 (2018). DOI 10.1145/3227609.3227670. URL http://doi.acm.org/10.1145/3227609.3227670
47. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: Data Engineering (ICDE), 2015 IEEE 31st International Conference on, pp. 651–662. IEEE (2015)
48. Wu, L., Liu, H.: Tracing fake-news footprints: Characterizing social media messages by how they propagate. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 637–645. ACM (2018)
49. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications **76**(4), 4801–4834 (2017)
50. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., Spangenberg, J.: Web and social media image forensics for news professionals. In: Tenth International AAAI Conference on Web and Social Media (2016)
51. Zandi, M., Mahmoudi-Aznaveh, A., Talebpour, A.: Iterative copy-move forgery detection based on a new interest point detector. IEEE Transactions on Information Forensics and Security **11**(11), 2499–2512 (2016)
52. Zollo, F., Novak, P.K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., Quattrociocchi, W.: Emotional dynamics in the age of misinformation. PloS one **10**(9), e0138,740 (2015)
53. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: A survey. ACM Computing Surveys (CSUR) **51**(2), 32 (2018)