A Corpus of Debunked and Verified User-Generated Videos

Purpose – As User-Generated Content (UGC) is entering the news cycle alongside content captured by news professionals, it is important to detect misleading content as early as possible and avoid disseminating it. This paper presents an annotated dataset of 380 user-generated videos, 200 debunked and 180 verified, along with 5,195 near-duplicate reposted versions of them, and a set of automatic verification experiments aimed to serve as a baseline for future comparisons.

Design/methodology/approach – The dataset was formed using a systematic process combining text search and near-duplicate video retrieval, followed by manual annotation using a set of journalism-inspired guidelines. Following the formation of the dataset, the automatic verification step was carried out using machine learning over a set of well-established features.

Findings – Analysis of the dataset shows distinctive patterns in the spread of verified versus debunked videos, and application of state-of-the-art machine learning models shows that the dataset poses a particularly challenging problem to automatic methods.

Research limitations/implications – Practical limitations constrained the current collection to three platforms: YouTube, Facebook, and Twitter. Furthermore, there exists a wealth of information that can be drawn from the dataset analysis, which goes beyond the constraints of a single paper. Extension to other platforms and further analysis will be the object of subsequent research.

Practical implications – The dataset analysis indicates directions for future automatic video verification algorithms, and the dataset itself provides a challenging benchmark.

Social implications – Having a carefully collected and labelled dataset of debunked and verified videos is an important resource both for developing effective disinformation-countering tools but also for supporting media literacy activities.

Originality/value – Besides its importance as a unique benchmark for research in automatic verification, the analysis also allows a glimpse into the dissemination patterns of UGC, and possible telltale differences between fake and real content.

Keywords Video verification, fake news, disinformation detection, user-generated content, social media, dataset

Paper type Research paper

1. Introduction

User-Generated Content (UGC), i.e. media content generated by non-professional bystanders during unfolding newsworthy events, has become an essential component of evolving news stories. The ubiquity of capturing devices means that it is very likely that bystanders may be capturing relevant content and sharing it through various Web and social media platforms. News professionals are pressed by competition to integrate such content in their stories, but verifying it first is essential to any news provider's reputation (Hermida & Thurman, 2008). Automatic and semi-automatic tools have the potential of considerably easing and speeding up the verification of user-generated content.

News content verification through automated means is a relatively young field, comprising a set of distinct disciplines, including rumour analysis (Zubiaga et al, 2018), multimedia forensics (Zampoglou et al, 2017), classification of social media content (Castillo et al, 2011), web mining and multimedia retrieval (Xie et al, 2011). A recent survey (Kumar & Shah, 2018) presented an analysis of known patterns of disinformation dissemination and approaches on automatic detection of false information.

Datasets are an important asset for understanding and addressing the problem of news content verification, and range from collections of tampered multimedia content and social media posts, to "rumours", i.e. cascades of unverified information. Carefully designed datasets may contribute both to better understanding the patterns of disinformation dissemination and to training and evaluating automatic detection systems.

This paper deals with User-Generated Video (UGV) verification, specifically with the effort to discern whether a suspect video conveys factual information or disinformation -in other words, for the sake of brevity, if the video is "real" or "fake". The paper presents the first large-scale video verification dataset, consisting of 380 videos and their 5,195 near-duplicates collected from YouTube (YT), Facebook (FB), and Twitter (TW), including a number of fake and real UGVs and numerous other versions of those videos that were consecutively posted online. The dataset is supplemented with 77,258 tweets that contain links to the dataset's videos. The dataset, named Fake Video Corpus 2018 (FVC-2018), which has been made publicly available[1], was gathered using a systematic process and can provide insights to the nature of disinformation, and the types of fake and real content circulating the Web. It is also aimed to serve as a benchmark for automatic content verification methods.

2. Related work

The area of multimedia verification consists of several fields of study, tackling various aspects of the problem from different viewpoints.

a. Multimedia forensics

A large part of related research concerns tampering detection and image/video forensics algorithms. Proposed algorithms attempt to detect and localize image modifications, either actively by embedding watermarks in multimedia content and monitoring their integrity (Dadkhah et al, 2014) (Botta et al, 2015), or passively by searching for telltale self-repetitions (Zandi et al, 2016) (Ferreira et al, 2016) or inconsistencies in the image. Such inconsistencies may appear in the pixel domain or the compressed domain depending on the specific process of tampering. A recent survey and evaluation of such algorithms can be found in (Zampoglou et al, 2017). Generally, such content-based approaches suffer from a number of issues that often render them inapplicable. One problem is their limited robustness with respect to image transformations. When the images or videos are recompressed or rescaled, as is often the case with social media uploads, the traces of the tampering tend to disappear (Zampoglou et al, 2017). Another limitation is that such approaches are only relevant in specific cases of disinformation. There are cases where a multimedia item is used to convey false information not by altering its content but by altering its context. One typical such approach is to reuse content from a past event and present it as if it is was captured during a current one. Another is to misrepresent the content, e.g. the location where it was taken or the identities of depicted

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

people. In such cases, an approach must be able to evaluate the context of the post (e.g. the profile of the uploader, the linguistic characteristics of the accompanying post, or the collective characteristics of all posts sharing the same item) rather than its actual content.

b. Automated fact checking

In automated fact checking (Hassan et al, 2015), 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. Furthermore, the automatic extraction of claims that can be cross-checked with a database is very difficult for audiovisual content. Thus, while it is a promising field, it is not currently suitable for UGV verification.

c. Rumour analysis

With the rise of social media, attention shifted to other aspects of verification. Twitter –and micro-blogging platforms in general- have attracted a lot of attention in the recent past. Several approaches operate at the "event" level, e.g. sets of tweets discussing one event or statement. The task of analysing a collection of social media posts around a claim is commonly referred to as *rumour detection*, defining rumour as a piece of information that may or may not be true (Zubiaga et al, 2018). In that definition, *rumour detection* refers to the process of gathering all posts related to a rumour. A classification or verification process can then be used to ascertain whether the rumour is true or not. Work by Castillo et al. (2011) was the earliest attempt in this category, presenting an algorithm to classify statements pertaining to events into "truthful" or "untruthful". Recent approaches (Vosoughi et al, 2017) attempt to develop methods for estimating the veracity of rumours by aggregating all posts disseminating them. The typical methodology of such methods is generally the same: a number of features are extracted from the tweet texts, the user profiles, and the internal structure of the topic (e.g. retweets), and a dataset of annotated rumours is used to train a classifier. A recent survey of approaches and datasets for rumour detection and classification can be found in (Zubiaga et al, 2018).

d. Tweet/post verification

There exist several verification approaches which aim to classify single posts, without taking into account other similar posts. This is an important distinction, since it ultimately concerns the speed at which an investigator can come to a conclusion about a piece of information. From the moment that the first post (tweet) appears making a claim, to the point where enough posts have been gathered into a "rumour", the time delay may be too long for news cycle standards. For that reason, having a system that can operate at the level of single posts is very useful. Such an approach was presented in (Gupta et al, 2013), where a set of features are extracted from the tweet text and the user profile and are used to classify tweets as truthful or not. A similar attempt (Wu et al, 2015) was used to classify microblog posts from the Sina Weibo platform. The classification of social multimedia by exploiting the associated tweet and user information was the aim of the "Verifying Multimedia Use" benchmarking task, which took place in MediaEval 2015 (Boididou et al, 2015) and 2016 (Boididou et al, 2016). A recent study (Boididou et al, 2018) compares the three top performing methods in this task.

e. Contextual video verification

The work presented here is aimed at video verification. With the exception of a body of works in video forensics which, as explained above, have several limitations in terms of applicability, there is one relevant recent work that attempts to tackle the problem using contextual

information (Papadopoulou et al, 2017). In this approach, a small dataset of YouTube videos, called Fake Video Corpus (FVC) was used to train and evaluate an automated classifier. The dataset contains around 104 videos annotated as “real” or “fake”, and the approach combines a classifier based on video and channel/user metadata features and a second classifier based on comment-based credibility features. This approach is limited in terms of dataset size and its results can only be treated as indicative.

The dataset presented in this paper, called the FVC-2018, builds upon the data of (Papadopoulou et al, 2017), however the methodology and scale are significantly different. Besides extending to a much larger number of cases, in this paper multiple copies of the same video are also collected from multiple platforms. This means that the total number of items in the FVC-2018 is an order of magnitude larger than that of FVC, and much more varied. More importantly, it potentially allows us to move beyond single-item methods, to approaches inspired by rumour detection (i.e. exploiting the presence of multiple items in each “case” to be verified to extract features from its collective properties and temporal evolution). This is partly related to the work of (Xie et al, 2011) where partial duplicates of news-related videos were gathered to analyze the dissemination of so-called “visual memes”, i.e. short video segments passed from one uploader to the other. While their work is in some aspects similar to the one presented here, it did not address the problem of fake videos. Furthermore, while (Xie et al, 2011) track all news-related content regardless of origin, this work deals with UGV verification specifically.

3. Methodology

a. Design and concepts

The FVC-2018 is aimed to serve both as a basis for analysis of the dissemination of UGV (*real* and *fake* videos), and as an evaluation benchmark for video verification systems. The definition of *fake* videos used here follows that of (Papadopoulou et al, 2017) and includes the following:

1. Staged videos where actors perform scripted actions under direction, falsely presented as authentic UGC captured during an event of interest.
2. Videos where contextual information is false (e.g. the claimed video location is wrong).
3. Past videos presented as being captured during unfolding events.
4. Videos of which the visual or audio content has been altered through editing.
5. Computer-generated Imagery (CGI) posing as real.

Real videos are videos that convey actual facts. For the formation of the dataset, this means that they need to have been verified first. Videos of which the veracity could not be confirmed with confidence were not included in the dataset.

There is a limited number of videos that can be collected like this, as the process is constrained to established cases of fake and real videos. However, when a newsworthy video is uploaded, and especially when it makes an unusual claim (regardless of its veracity), it tends to get further disseminated by users. That is, people tend to share and re-upload the content, usually with no mention of the original source. These versions are often slightly altered, not only because of the downloading and recompression, but also by various forms of editing, e.g. by adding highlights, slow motion, commentary, or by changing the audio. Thus, each newsworthy UGV tends to be followed by a cascade of other versions, and the overall social media activity around them. All

this information can be critical for verification, as it can be used both by human investigators and by automatic contextual analysis approaches.
In order to collect alternate versions from videos, search methods combined with near-duplicate detection tools were used. Near-duplicate retrieval is the task of locating (within a given collection) all videos that visually resemble a given query video. While "visual resemblance" is not a strictly defined term, most near-duplicate retrieval tasks focus on the retrieval of different versions of the original content, e.g. following editing, post-processing, cropping, etc. Near-duplicate video detection algorithms have achieved significant progress in the recent past (Kordopatis-Zilos et al, 2017), and are mature enough for real-world application.

b. Dataset collection

The Fake Video Corpus dataset was used as the initial basis of this work. The FVC contains 104 videos, of which 55 were annotated as *fake* and 49 as *real*. The authors of (Papadopoulou et al, 2017) created the dataset over an extended period of time (2016-2017) in cooperation with media experts from the InVID[2] project to serve as a representative collection of past fake videos, and was manually extended to also contain a number of real news-related UGVs. The first step was to extend it with more cases, both fake and real ones. Between April and July 2017, the dataset was manually extended and reached 117 fake videos and 110 real videos. However, manually gathering news-related UGVs is not a straightforward task. In order to further extend the dataset, one additional valuable source was the Context Aggregation and Analysis service[3], which was developed within the InVID project as a tool for video verification. The service, being one of the few publicly available tools for video metadata analysis, generally attracts traffic from verification experts who submit suspicious videos for verification. All videos submitted to the service between November 2017 and January 2018 resulted in an initial pool of approximately 1600 videos. This set was filtered to remove non-UGC and other irrelevant content, and consecutively, was annotated as *real* or *fake*. For this initial annotation, debunking sites such as *snopes.com* were used in addition to –especially for *real* content- the general consensus that reliable news sources publish factual content. Furthermore, *snopes.com* and other debunking sites were consulted in order to collect more debunked *fake* videos.

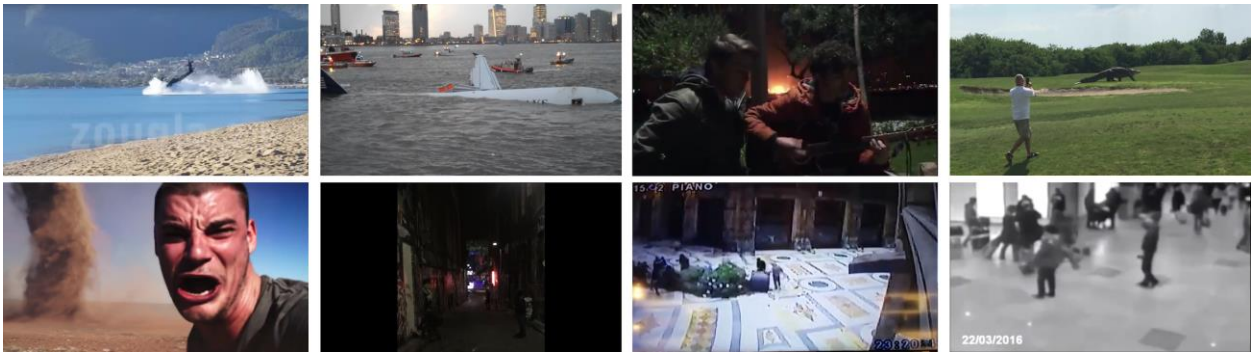

Figure 1: Indicative cases of real and fake User-Generated Videos. Top: four real videos. a) A Greek army helicopter crashing into the sea in front of beach; b) US Airways Flight 1549 ditched in the Hudson River; c) A group of musicians playing in an Istanbul park while bombs explode outside the stadium

[2] http://www.invid-project.eu
[3] http://caa.iti.gr/

behind them; d) A giant alligator crossing a Florida golf course. Bottom: four fake videos. a) "A man taking a selfie with a tornado" -CGI; b) "The artist Banksy caught in action" -staged; c) "Muslims destroying a Christmas tree in Italy" -out of context, there is no indication that the men are Muslim; d) "Bomb attack on Brussels airport" -out of context, the footage is from Moscow Domodedovo airport a few years back.

Using the above process, an initial set of 380 videos was formed, of which 200 were annotated as *fake* and 180 as *real*. Figure 1 presents some indicative cases. This collection is in itself a useful dataset for the field and the result of a significant amount of manual effort and accumulated knowledge. However, a major contribution of this work is the decision to move beyond isolated videos and try to explore how such content is disseminated and reposted through time in various platforms.

To this end, the following approach was followed in order to end up with a collection of videos that would be highly likely to contain several versions of the query video:

1. For each video in the original set, extract the video title
2. Reformulate the title of the video in a more general form (called the "event title"). For example, a video with title "Video Tornado IRMA en Florida EEUU Video impactante" was assigned the event title "Tornado IRMA at Florida"
3. Translate the event title from English into four major languages[4]: Russian, Arabic, French, and German using Google Translate
4. Use the video title, event title, and the four translations as separate queries to the three target platforms: YouTube, Facebook, Twitter. Group all returned videos in a common pool
5. Use the near-duplicate retrieval algorithm of (Kordopatis-Zilos et al, 2017) to search within this pool for near-duplicates of the video
6. Apply a manual confirmation step to remove any erroneous results of the method and only retain actual near-duplicates.

Using the above process, the collected videos were organized into "video cascades". The term refers to a set of videos, starting with a first posted video and including all its near-duplicates temporally ordered by publication time.

Methodologically, two further steps were applied to extend and refine the dataset. The first was to submit the URL of the first video of each cascade to Twitter search, and collect all tweets sharing the video as a link. This is a different type of item than the rest of the videos, since it is a link pointing to a video in the cascade accompanied by some text. However, the type of Twitter traffic that a video attracts can be a useful indicator of its credibility. The second was to scan the dataset and separate between "fake" items that reproduce the falsehood from ones that debunk it or use it for entertainment purposes, and correspondingly for "real" items. This is important because such videos should not be taken into account by certain analysis tasks.

c. Verification algorithm

One of the potential uses of the FVC-2018 dataset is to train and evaluate automatic verification algorithms. As a set of baseline results, such an algorithm (Papadopoulou et al, 2017) was applied to the dataset, and its performance on the data was evaluated.

[4] These languages were selected after preliminary tests indicated that near-duplicate videos appear with increased frequency in these languages.

The algorithm builds classification models over two sets of features (Table 1). The first set is based on the video metadata, and specifically linguistic features extracted from the video description text and statistics extracted from the video channel. These are concatenated to form a feature vector. The second feature set is based on the comments under the video, and the descriptor is extracted using a two-level approach. First, a set of features is extracted from each individual comment, as shown in Table 1. Then the credibility of each comment is independently evaluated using a pretrained model (Boididou et al, 2018). While the model of (Boididou et al, 2018) was trained on a large set of tweets using their linguistic features, it can similarly be applied to video comments. The classifier returns a credibility score in the range [0, 1]. By accumulating the scores of all video comments in a single 10-bin histogram, a vector of 10 variables is produced per video and used for the comment credibility classifier.

Table 1: Overview of video metadata and comment credibility features

Video metadata features	**Comment credibility features**
From channel description	01 Text length
01 Channel view count	02 Number of words
02 Channel comment count	03-04 Contains question/exclamation mark (Boolean)
03 Channel subscriber count	05-06 Contains happy/sad emoticon (Boolean)
04 Channel video count	07-09 Contains 1st/2nd/3rd person pronoun (Boolean)
From video description	10 Number of uppercase characters
05 Text length	11-12 Number of positive/negative sentiment words
06 Number of words	13 Number of slang words
07-08 Contains question/exclamation mark (Boolean)	14-15 Has ':' symbol/'please' (Boolean)
09-10 Contains 1st/3rd person pronoun (Boolean)	16-17 Number of question/exclamation marks
11 Number of uppercase characters	18 Readability score
12-13 Number of positive/negative sentiment words	
14 Number of slang words	
15 Has ':' symbol (Boolean)	
16-17 Number of question/exclamation marks	

In (Papadopoulou et al, 2017), metadata and comment descriptor vectors were used to train Support Vector Machine (SVM) classifiers. In the verification methodology used here, the same process was used, with the addition of two model variants: a) a concatenation of the two feature sets; b) the agreement-based approach of (Boididou et al, 2018) was used, where, following an initial classification of all videos using the two classifiers, the cases where the two classifiers agree are kept. The rest are re-classified using a concatenated feature, either trained on the original training set, or on a new training set consisting of the original plus the agreed-upon videos. Such approaches have been shown to increase classifier performance and were incorporated in the verification results presented with the dataset.

The dataset presented here contains videos from three different platforms. This presents a challenge in the form of unifying the descriptors and treating each video indistinguishably regardless of platform. The main issue with this is the fact that channel description features are not available for Facebook videos. Thus, in all experiments run in this dataset, whenever videos from all platforms are used, such features are not included. This may lead to a degradation of performance. Thus, for comparison, experiments using only the YouTube videos of the dataset were also run, to evaluate the potential of platform-specific models.

4. Results

a. *Dataset overview*

The initial, manual collection of videos resulted in 200 unique *fake* videos and 180 *real* unique videos. While in many cases these initial, manually collected videos were confirmed to be the first version of the video to be posted, there was no way to confirm this in every case prior to the application of the near-duplicate retrieval approach described in Section 4.a. Following that step, the dataset was extended with 3,729 additional *fake* videos and 2,283 *real* videos that partly or fully reproduced the submitted video, published on YouTube, Facebook, or Twitter, and covering a period between April 2006 and June 2018. Overall, 172 *fake* and 148 *real* videos had at least one near-duplicate in at least one video platform, while for the rest zero near-duplicates were found. One first noteworthy observation is that the number of collected fake videos is much larger than the corresponding number of real videos. This suggests that fake videos tend to be reproduced more, either as repeating acts of disinformation or in videos analyzing, parodying, or debunking the video. This is in line with the observations of Vosoughi et al (2017), who, while analyzing the distribution of Twitter rumours, also observed a higher rate of reproduction for misleading content.

As described in the “Methodology” section, the next step was to manually study the dataset and categorize the near-duplicates based on their intent. This led to five categories of near-duplicates of fake videos and four categories of near-duplicates of real videos.

The categories for near-duplicates of fake videos are: a) *Fake/Fake*: those that reproduce the same false claims; b) *Fake/Uncertain*: those that express doubts on the veracity of the claim; c) *Fake/Debunk*: those that attempt to debunk the original claim; d) *Fake/Parody*: those that use the content for fun/entertainment; e) *Fake/Real*: those that contain the earlier, original source from which the fake was made. For near-duplicates of real videos, the corresponding categories are: a) *Real/Real*: those that reproduce the same factual claims b) *Real/Uncertain,* those that express doubts on the veracity of the claim; c) *Real/Debunk*: those that attempt to debunk their claims as false; d) *Real/Parody*: those that use the content for fun/entertainment. A special category concerns videos labeled *Real/Private* and *Fake/Private*, which describes Facebook videos that were relevant to the dataset but were published by individual users and thus could not be accessed through the API in order to extract their context. These were left out of the analysis entirely. Table 2 shows the number of videos that corresponded to each category and each platform. The column labeled “Total” corresponds to all videos, and does not include the twitter posts that share the video, which are counted separately.

The most time-consuming part of the process was forming and annotating the initial set of 380 videos, which was aided by the fact that most of the videos have already been discussed online. Following that, the annotation process of the near-duplicates was generally straightforward and

fast, due to the high degree of content repetition between near-duplicates . As a result, after the authors' team collectively concluded on the appropriate tag for each case's initial video, the individual videos could be correctly and swiftly annotated by a single person. To make the task more manageable, the annotation was assigned to two annotators, each of whom was assigned a different part of the videos to annotate. The annotation required in total roughly 20 hours for the YouTube videos and 20 hours for the Facebook videos, while the annotation of tweets sharing the videos took roughly 180 hours.

Table 2: Types of near-duplicate videos collected. Private videos are not included in the totals.

Fake videos						**Real videos**					
	YT	FB	TW	Total	TW shares		YT	FB	TW	Total	TW shares
Initial	189	11	0	200	-	Initial	158	22	0	180	-
Fake	1,675	928	113	2,716	44,898	Real	993	901	16	1,910	28,263
Private	-	*467*	-	*467*	-	Private	-	*350*	-	*350*	-
Uncertain	207	122	10	339	3,897	Uncertain	0	1	0	1	30
Debunk	68	19	0	87	170	Debunk	2	0	0	2	0
Parody	43	2	1	46	0	Parody	14	6	0	20	0
Real	22	51	1	74	0						
Total	**2,204**	**1,133**	**125**	**3,462**	**48,965**	**Total**	**1,167**	**930**	**16**	**2,113**	**28,293**

While all types of videos were retained in the FVC-2018 dataset for potential future analysis, the ones considered relevant to the analysis presented here 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 and ignore the rest.

Overall, the scale of the FVC-2018 dataset is comparable to existing datasets for rumour verification. In comparison, the dataset of Gupta et al (2013) 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 dataset of Vosoughi et al (2017) contains 209 rumours with – on average- more than 3,000 tweets each, of which the collection was carried out automatically in order to reach this scale. One important distinction from 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, which makes their collection a challenging problem.

b. *Video and description characteristics*

Certain aspects of the accumulated data can prove useful for the analysis of the dataset and are worth highlighting. These concern the videos themselves, their accompanying text (post, video description) and the posting account. In analyzing these characteristics a first approach is to

compare the distribution of various features for fake and real videos. In the rest of this section, when comparing feature distributions we present either the mean or the median, depending on whether the variable follows a normal distribution or not. To evaluate the statistical significance of the corresponding differences, we compare the means using Welch's t-test or the Mann–Whitney–Wilcoxon (MWW) test respectively and report the associated p-values.

One interesting feature is the difference in video durations. Real videos have an average duration of 149 seconds if we only take the initial videos into account, and 124 seconds if we include the edited near-duplicates In contrast, initial fake videos have a much smaller average duration of 92 seconds ($p<10^{-3}$), and their near-duplicates have an average duration of 77 ($p<10^{-3}$) seconds. While this feature itself is not sufficient to classify a video as fake or real, it is interesting to note that fake videos tend to be significantly shorter. Another interesting distinction is the age of the channel/account posting the video. For real videos, the YouTube channel median age at the time of posting is 811 days prior to the video publication, while for fake videos this value is 425 ($p<10^{-3}$). For Twitter, these values are 2,325 and 473 days ($p=10^{-3}$), and for Twitter shares the corresponding values are 1,297 and 1,127 respectively. In the latter case, while the difference does not seem large, given the large sample size it is still statistically significant ($p<10^{-3}$). No such information was available for Facebook. Overall, fake videos tend to be posted by "younger" YouTube and Twitter accounts compared to real videos. With respect to the channel/account, it is also interesting to observe that, for real videos, the median YouTube channel subscriber count is 349 users, while for Twitter the median follower count is 163,325. The latter value is particularly high due to the fact that only a small number (16) of well-established Twitter accounts with many followers were found to have re-uploaded the content as a native twitter video. This is contrasted to the median number of followers of the Twitter accounts which shared the video as a link, which was 333. For the Fake videos the median follower counts are much lower: 98 ($p<10^{-3}$) and 2,855 ($p<10^{-3}$) for YouTube and Twitter respectively. For Twitter shares the median follower count was 297 which, while closer to the one for real videos, is still significantly lower ($p<10^{-3}$).

Other interesting conclusions stem from linguistic analysis of the text that accompanies the videos (video description or post text depending on the platform). First, language was automatically detected for all posts using the Python langdetect[5] library. For real videos, the relatively most frequent language in the texts is English (YT: 63%, FB: 41%, TW: 75%, TW shares: 62%). For fake videos the corresponding values are noticeably lower, although still high (YT: 38%, FB: 28%, TW: 43%, TW shares: 52%). A sizeable number of posts/descriptions did not contain enough text for automatic recognition, although that number was generally smaller for real videos (YT: 13%, FB: 48%, TW: 0%, TW shares: 5%) than for fake ones (YT: 28%, FB: 51%, TW: 0%, TW shares: 4%). Other languages encountered in the set included Russian, Spanish, Arabic, German, Catalan, Japanese, and Portuguese. With the exception of Russian fake Twitter videos which are strikingly high (28%), these languages appear at a frequency of less than 6% in each category. A number of features were calculated from this text, namely Polarity, Subjectivity, Flesh reading ease (Kincaid et al, 1975) and the Coleman-Liau index (Coleman and Liau, 1975). Polarity and Subjectivity were calculated using the Python TextBlob library[6] while the other two were calculated using the Python textstat[7] library. No noticeable

[5] https://pypi.org/project/langdetect/
[6] http://textblob.readthedocs.io/en/dev/

differences were found between fake and real videos. Despite the common assumption that fake posts have distinctive linguistic qualities, e.g. stronger sentiment and poorer language (Castillo et al, 2011), no such pattern was found in our study.

c. *Temporal distribution*

An important aspect of the FVC-2018 dataset is the temporal distribution of the near-duplicates, and their relative importance in terms of popularity and user attention. To explore this, a timeline was created (Figure 2), showing all near-duplicates per cascade. This shows how the near-duplicates of real and fake videos are distributed in the dataset. Each line corresponds to one original video and its near-duplicates (i.e. a cascade). The horizontal axis corresponds to the time (log-scale) between the posting of the initial version and its near-duplicates. In principle, each dot corresponds to a near-duplicate being posted at that time, and the color of the dots corresponds to the platform (red: YouTube; blue: Facebook; green: Twitter; light blue: sharing the original video link as a tweet). The videos are sorted from the most duplicated ones at the top, to the least duplicated ones at the bottom.

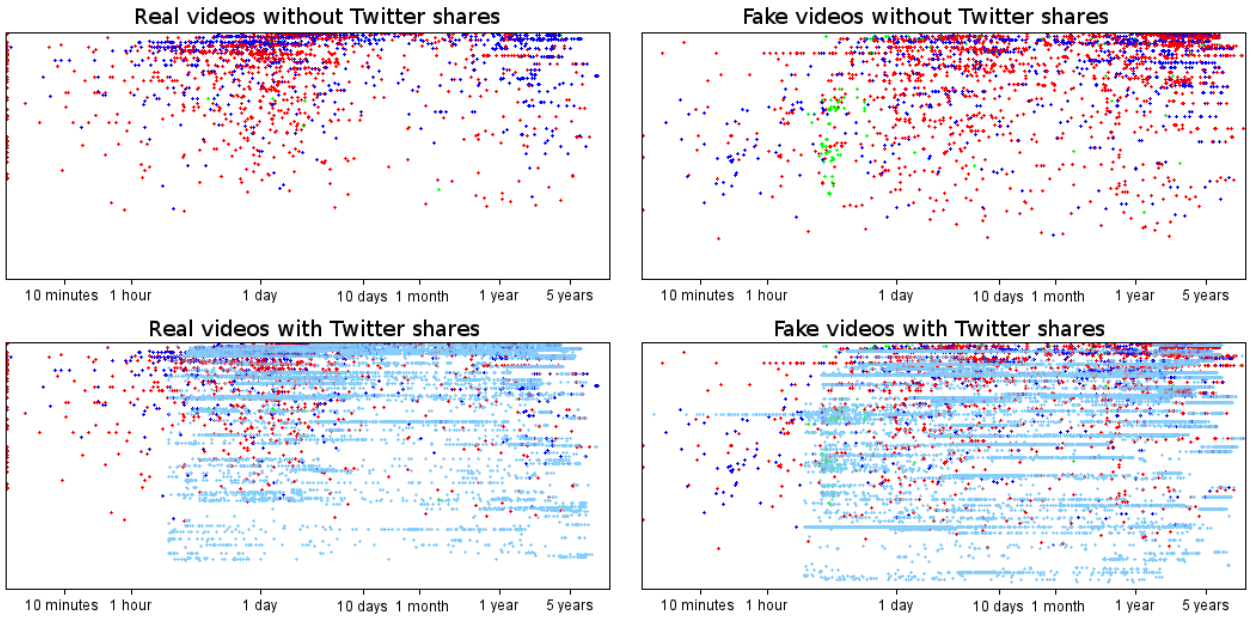

Figure 2: Temporal distribution of video near-duplicates. Red: YouTube; Blue: Facebook; Green: Twitter; Light blue: posting on Twitter (as link) of the first video in the cascade.

The time range of the dataset reaches a maximum at about 10 years between the first posting of a video and its most recent near-duplicate. The difference in the temporal distribution of the fake videos compared to that of real ones is conspicuous. There are relatively few near-duplicates of real videos posted on YouTube after 10 days from the original post, and the same pattern also holds for Twitter shares. Instead, for fake videos, near-duplicates are posted at a much higher rate for a much longer interval. This discrepancy is also reflected in the fact that the median time difference between the initial video and its near-duplicates is much higher for fake videos than real ones on YouTube and Facebook. While for real videos the median

[7] https://pypi.org/project/textstat/

temporal distance is 1 and 3 days respectively, for fake videos the corresponding values are 62 ($p<10^{-3}$) and 148 ($p<10^{-3}$). For Twitter videos the values are comparable, 1 and 0 days for real and fake videos respectively, although the difference is still significant ($p=3\times10^{-2}$), but this concerns only a few items. For videos tweeted as links, the median distance is 6 days for real videos and 27 days for fake videos ($p<10^{-3}$).

d. *Video categories*

Another interesting feature of the collected videos is the category assigned to them by their uploader. YouTube and Facebook both have a “Category” tag allowing the user to categorize the video. Figure 3 shows the distribution of category tags for fake and real videos on YouTube and Facebook. Twitter does not offer a corresponding feature and was thus not considered. One may observe clear differences between fake and real videos on both platforms. Real videos on Facebook tend to be categorized as “News” more frequently than fake ones, where the “Entertainment” and “Music” categories are more prominent. Similarly for YouTube, categories “Video blogging” and “Comedy” are more frequent for fake videos. Taken alone, this distinction is not enough to identify disinformation. However, it may offer a useful verification signal in some cases.

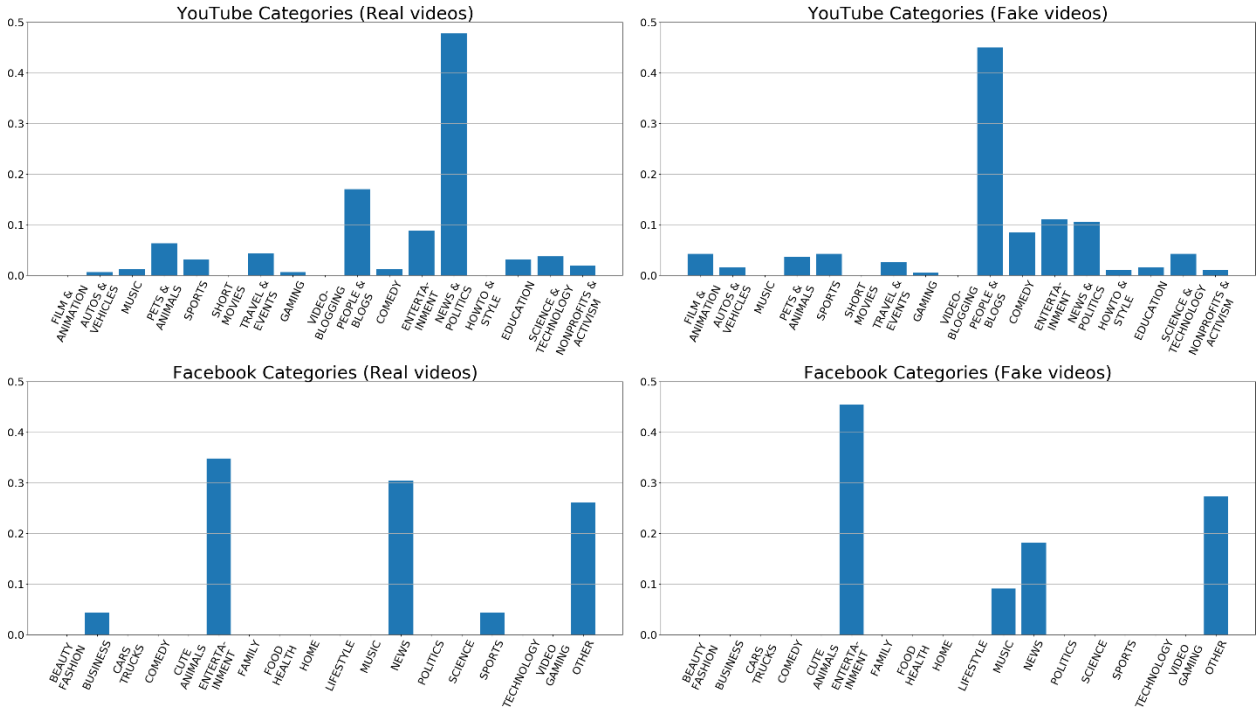

Figure 3: Video category distribution for YouTube and Facebook.

e. *Video comments*

Another noteworthy component of the dataset is the number and distribution of user comments in the videos. Comments can provide a sort of “wisdom of the crowd”, which may assist investigators in identifying inconsistencies that expose fake UGVs, or by providing clues that strengthen the claims made by a video. Previous research (Papadopoulou et al, 2017) has also highlighted the importance of user comments for automatic verification. It is therefore important

to study the comment distribution in FVC-2018. Overall, the dataset contains 491,636 comments on *fake* YouTube videos, and 433,139 comments on *real* ones, 105,814 and 86,326 respectively for Facebook videos, and 561 and 215 for Twitter videos (in this case, we treat replies as comments). A significant percentage of these, especially for YouTube, is found in the first video of each cascade (YouTube: 81% and 69% for fake and real videos; Facebook: 22% and 9% respectively). Figure 4 presents the cumulative average number of comments over time per video for the three video platforms.

There are some features that stand out. One is the difference in the number of comments between platforms, with YouTube attracting significantly more than the others. The second is the difference between the number of comments on fake and real videos on YouTube, with real ones attracting much more comments. A third observation is the steep increase in the number of YouTube comments in real videos, between 12 hours and 10 days after the video is posted, which consecutively tapers off. In contrast, fake videos maintain a steadier rate of accumulation, which, especially after one year from the posting, ends up relatively steeper than for real videos.

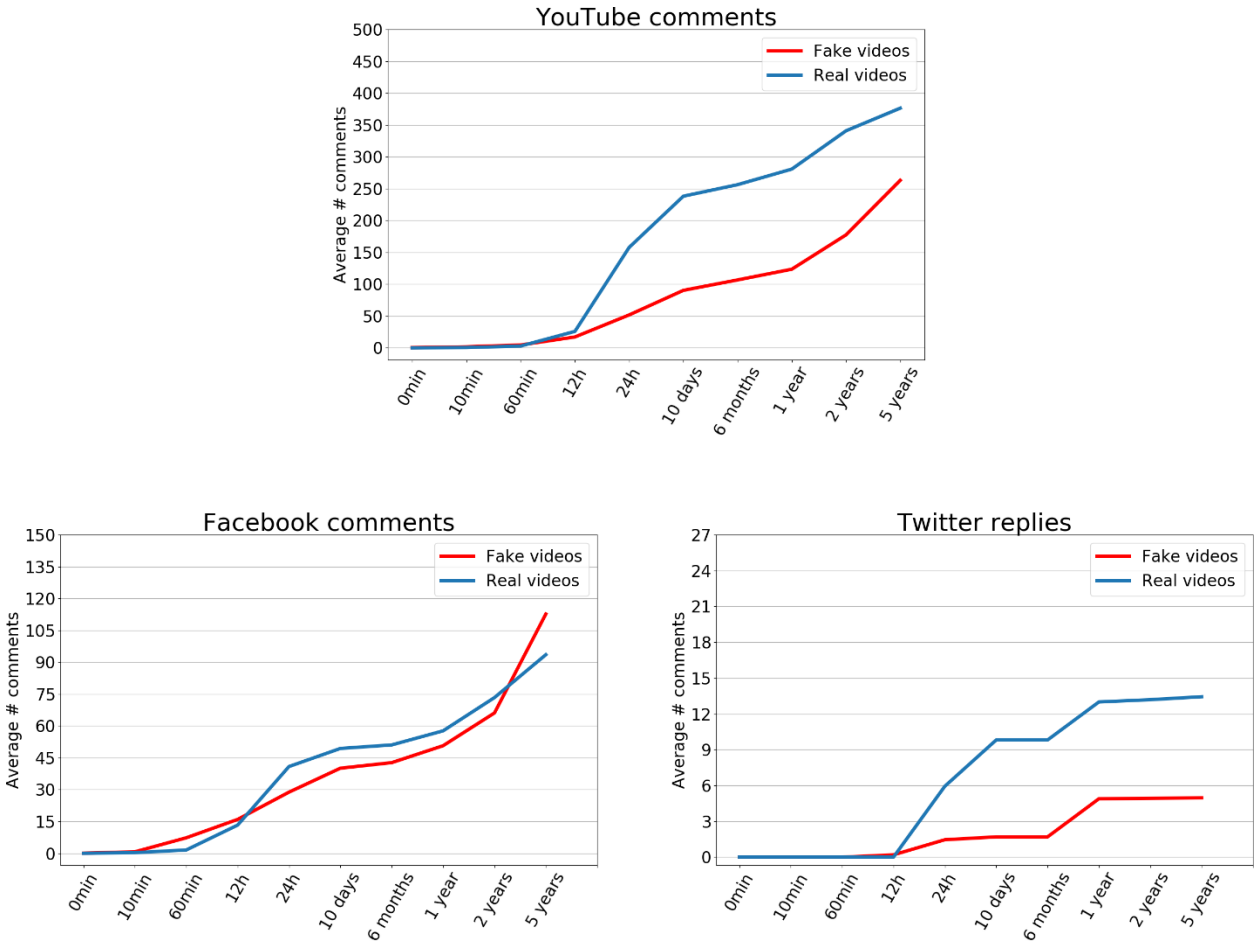

Figure 4: Timeline view of video comment accumulation.

f. *Automatic verification*

This section provides a set of experimental results using an existing state-of-the-art automatic verification approach, which can serve as a benchmark for future methods and provides an estimate of the level of challenge that the dataset poses to automatic methods.

The video metadata and comment credibility features described in Section 3.b were extracted from all videos where this was possible, and the approach of (Papadopoulou et al, 2017) was applied for verification. Two separate sets of experiments were run: The first set used only the oldest video from each cascade, resembling the setting of (Papadopoulou et al, 2017). The second set used all videos in the dataset. In both cases, one run considered only YouTube videos, in order to exploit all possible features, and the other considered videos from all platforms, with reduced features in order to create uniform descriptors. Since there are no first videos from Twitter, the first video per cascade run does not include Twitter videos.

The evaluations were run using 10-fold cross-validation. For the experiments using all near-duplicates, the cross-validation was cascade-based, i.e. all videos from the same cascade were put into the same fold. This ensures that there is no leakage of information between the training and test sets. The results of the runs are presented in Table 3.

Table 3: Automatic verification results for the dataset of (Papadopoulou et al, 2017), the first video per cascade, and the entire FVC-2018. In all labels, "Prec": Precision, "Rec": Recall, "F1": F1-score.

	(Papadopoulou et al, 2017)	**First video per cascade (YT only)**	**First video per cascade (YT + FB)**	**All videos in the cascade (YT only)**	**All videos in the cascade (YT + FB + TW)**
Comment credibility	Prec.: 0.88 Rec.: 0.74 F1: 0.79	Prec.: 0.91 Rec.: 0.53 F1: 0.67	Prec.: 0.97 Rec.: 0.52 F1: 0.68	Prec.: 0.96 Rec.: 0.64 F1: 0.77	Prec.: 0.94 Rec.: 0.60 F1: 0.73
Video metadata	Prec.: 0.88 Rec.: 0.79 F1: 0.82	Prec.: 0.87 Rec.: 0.59 F1: 0.70	Prec.: 0.87 Rec.: 0.58 F1: 0.70	Prec.: 0.95 Rec.: 0.69 F1: 0.80	Prec.: 0.95 Rec.: 0.60 F1: 0.74
Concat.	Prec.: 0.88 Rec.: 0.82 F1: 0.85	Prec.: 0.79 Rec.: 0.61 F1: 0.69	Prec.: 0.77 Rec.: 0.60 F1: 0.67	Prec.: 0.92 Rec.: 0.70 F1: 0.79	Prec.: 0.87 Rec.: 0.64 F1: 0.74
Agreement-based	Prec.: 0.84 Rec: 0.88 F1: 0.86	Prec.: 0.58 Rec.: 0.93 F1: 0.71	Prec.: 0.53 Rec.: 0.98 F1: 0.70	Prec.: 0.70 Rec.: 0.96 F1: 0.80	Prec.: 0.61 Rec.: 0.96 F1: 0.74
Agreement-based with retraining	Prec.: 0.77 Rec.: 0.86 F1: 0.81	Prec.: 0.57 Rec.: 0.92 F1: 0.70	Prec.: 0.54 Rec.: 0.98 F1: 0.69	Prec.: 0.69 Rec.: 0.96 F1: 0.80	Prec.: 0.60 Rec.: 0.97 F1: 0.74
Ideal fusion	Prec.: 1.00 Rec.: 0.83 F1: 0.90	Prec.: 0.64 Rec.: 0.99 F1: 0.79	Prec.: 0.56 Rec.: 0.99 F1: 0.71	Prec.: 0.73 Rec.: 0.99 F1: 0.84	Prec.: 0.64 Rec.: 0.99 F1: 0.78

The rows correspond to different feature sets. The first two correspond to the basic features described in Section 3.b, and the third to their concatenation. "Agreement-based" refers to the practice of separately using the comment credibility and video metadata models, keeping the videos for which the two classifications agree, and re-classifying the rest using the concatenated feature vector. "Agreement-based with retraining" refers to a similar approach, the difference

being that the concatenation-based classifier is retrained using the videos for which the classifiers agreed, thus providing some additional adaptation to the dataset. For every run, the table shows the Precision, Recall, and F1-score. Finally, “ideal fusion” is a theoretical result which takes the outputs of the comment credibility and video metadata classifier, and assumes there exists a perfect fusion system that knows which one is correct on every case. Thus, it correctly classifies a video if at least one of the two results is correct. This provides an estimate of the maximum performance possible if the system had access to the best possible fusion approach.

The evaluation metrics show a degradation of performance on the new dataset compared to the earlier experiments of (Papadopoulou et al, 2017), both in the case of only using the first posted video in each cascade and when using all the videos in the dataset. When looking at the F1 scores, results are significantly lower than the first column in all cases. Furthermore, removing the channel-based features in order to merge Facebook, Twitter and YouTube videos leads to significantly reduced performance, both when using only the first video of each cascade and for the entire cascade. Ideal fusion between the two features does increase performance, but not at the levels of (Papadopoulou et al, 2017)

5. Discussion

a. Video and channel characteristics

The aim of this work is to provide insights into the dissemination patterns of fake and real UGVs, in order to assist verification. The results of Section 4 give ground for several such observations. A first set of observations concern the video and associated text. To the extent that we consider the dataset to be representative of the UGV circulating the Web, there are certain patterns that distinguish fake from real videos. One is the length of the video itself, and another is the fact that outlets sharing fake videos tend to be younger and have fewer followers. Another characteristic is the lack of distinctive differences in linguistic terms between fake and real videos. While such features have in the past been used in automatic verification systems (Castillo et al., 2011) (Wu et al, 2015) (Boididou et al, 2018), it seems that on this dataset these features do not have strong distinctive power. However, that does not mean that such features are useless, since machine learning algorithms may identify complex correlations among them and utilize them for improving classification performance. The same applies to the statistical differences between content categories for fake and real videos on YouTube and Facebook. While these differences cannot be directly used as a criterion by a human investigator, they may prove useful if combined with other distinguishing features.

Another observation pertains to the large time range that the circulation of a given video may span. In contrast to rumour verification, where generally the duration of rumours is in the scale of hours (Vosoughi et al, 2017), the dissemination of fake videos may cover an entire decade. From a verification perspective, this means that a large bulk of fake videos posted at any given time are actually old fakes that have probably already been debunked. This implies that investigators and the general public could dismiss the majority of misleading content if they had the means to easily match them with their previous, debunked versions.

b. *Comment distribution*

With respect to comment distribution, the overall difference in the average number of comments between fake and real videos can be attributed, at least in part, to the relatively smaller number of near-duplicates for real videos. Indeed, the *fake* set contains many near-duplicates that attracted relatively little user discussion. On the other hand, *real* videos are generally not reproduced as much, rather containing a small number of videos with much higher engagement. The high number of near-duplicates (often without comments) for fake videos highlights the fact that only a few fake video reposts have a noticeable disinformation impact.

One explanation for the steady increase of comments in fake videos through time is the tendency of users to link to old fake videos in social media platforms in the context of unfolding news events. This is clearly observed in Figure 3, in which tweets sharing the video links to fake videos appear consistently throughout the decade that the dataset covers. When a video is uploaded with a new date, perhaps slightly edited, it might convincingly appear as a new event and might even mislead an experienced investigator. On the other hand, the practice of simply posting a link to an old fake video is a cruder form of misinformation that is unlikely to fool a professional. It nonetheless seems to be able to misguide part of the public and reinvigorate traffic in an old video, even one that is several years old. Thus, *fake* videos tend to remain engaging for longer periods, compared to *real* newsworthy videos which tend to exhaust their engagement with the passing of time both in terms of comment activity in old uploads and with respect to the possibility of being re-uploaded.

c. *Automatic verification*

The observation that user/channel features were important for classification is interesting in the sense that it comes in contrast to the observations of (Gupta et al, 2013), where user features (on Twitter) led to a degradation of performance. Overall, the experiments imply that existing approaches building on supervised learning and fusion models may be currently inadequate to deal with the complexity of the problem. This is in contrast to recent works in the related field of rumour/tweet verification, where such methods seem to lead to high classification accuracy (Gupta et al, 2013) (Vosoughi et al, 2017) (Boididou et al, 2018).

Another important consideration derives from the observation that the experiments using videos from all platforms by removing the channel features led to significant degradation in accuracy. This highlights the importance of platform-specific models for verification. If videos from multiple sources are to be included in an analysis, designing dedicated models for them seems to be necessary in order to achieve good accuracy.

6. Conclusions

This work presented a novel annotated dataset of debunked and verified videos (termed *fake* and *real* respectively) collected from three platforms, and supplemented by the collection of Twitter posts that share the links to them, and organized into cascades. The dataset, named Fake Video Corpus 2018 (FVC-2018) is also accompanied by a set of experimental results using standard supervised learning and different fusion schemes. Besides its value as a benchmark for future approaches, a third novel contribution of this work is the analysis of the differences in the characteristics and evolution of real and fake video cascades over time with respect to the appearance of near-duplicates, the accumulation of comments and the distribution of various features.

The implications of this work for future research are manifold: The data gathering methodology itself followed a novel protocol, exploiting advances in near-duplicate video retrieval to move from isolated videos into cascades. Furthermore, analysis of the collected data showed certain interesting patterns in real and fake videos, which could be exploited by human investigators and verification algorithms. Also, with respect to automatic verification, while the dataset proved to be challenging for the algorithm used, the increased F1-score for the theoretical ideal fusion classifier over all YouTube videos shows that there could be potential for a fusion scheme to benefit from the relative complementarity of video metadata and comment credibility features.

The observations made on the distribution of fake videos in particular could also be of immediate practical importance. It was observed that videos keep attracting comments for a very long time after they are posted, as a result of being reposted in social media. In raising awareness against disinformation, flagging such old fakes on the platforms where they have been published could be an easy step that could assist users in identifying them and remove one significant source of disinformation. Also, given that a large proportion of misleading videos posted at any given time are actually near-duplicates of past, debunked videos, identifying them as such would greatly reduce the amount of disinformation circulated. With recent advances in near-duplicate detection, this could be a feasible task. Platforms such as YouTube already apply near-duplicate detection to detect copyright infringement. If the platforms would be willing to open these functionalities to third parties in order to run their own searches, this could empower investigators and the general public to timely identify and dismiss such fakes.

The current work has certain limitations which leave open possibilities for future research. One is the choice of video sharing platforms. Videos were collected from YouTube, Facebook, and Twitter. Other platforms, such as Instagram, could also be studied in the future. Also, the current work did not take into account another form of information sharing, namely messenger applications such as Viber and WhatsApp. While it is difficult to automatically collect information exchanged through such applications, including cascades from such sources would provide new important insights on the dissemination of disinformation. Another limitation is the mixing of established news outlets sharing real videos, with independent users sharing them. A more refined annotation step could be carried out on the dataset to distinguish between the two.

Furthermore, the limitations of the automatic classification method may be surpassed by taking advantage of the cascade structure provided by the dataset. This would be similar to rumour detection algorithms which exploit not only the individual features of posts, but the way they are disseminated and distributed within the cascade. While tools such as the InVID plugin (Teyssou et al, 2017) already provide relevant verification functionalities for isolated videos, research in cascade analysis could provide tools with more empowering functionalities, and the dataset presented here is aimed as a suitable evaluation benchmark for such a challenge.

Acknowledgement

This work has been supported by the InVID project, partially funded by the European Commission under contract no. H2020-687786.

References

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M. and Kompatsiaris, Y., 2015, September. Verifying Multimedia Use at MediaEval 2015. In *MediaEval*.

Boididou C., Papadopoulos S., Dang-Nguyen D., Boato G., Riegler M., Middleton SE., Petlund A., Kompatsiaris Y., 2016, Verifying Multimedia Use at MediaEval 2016. In *MediaEval*.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O. and Kompatsiaris, Y., 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, *7*(1), pp.71-86.

Botta, M., Cavagnino, D. and Pomponiu, V., 2015. Fragile watermarking using Karhunen–Loève transform: the KLT-F approach. *Soft Computing*, *19*(7), pp.1905-1919.

Coleman, M. and Liau, T. L. (1975); A computer readability formula designed for machine scoring, Journal of Applied Psychology, Vol. 60, pp. 283–284

Castillo, C., Mendoza, M. and Poblete, B., 2011, March. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web* (pp. 675-684). ACM.

Dadkhah, S., Manaf, A.A., Hori, Y., Hassanien, A.E. and Sadeghi, S., 2014. An effective SVD-based image tampering detection and self-recovery using active watermarking. *Signal Processing: Image Communication*, *29*(10), pp.1197-1210.

Ferreira, A., Felipussi, S.C., Alfaro, C., Fonseca, P., Vargas-Munoz, J.E., dos Santos, J.A. and Rocha, A., 2016. Behavior Knowledge Space-Based Fusion for Copy–Move Forgery Detection. *IEEE Transactions On Image Processing*, *25*(10), pp.4729-4742.

Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A., 2013, May. 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.

Hassan N., Adair B., Hamilton J., Li C., Tremayne M., Yang J., Yu C. (2015) The quest to automate fact-checking. *Proc. of the 2015 Computation and Journalism Symposium*, pp. 1-5

Hermida, A., and Thurman, N. (2008). A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites. *Journalism practice*, 2(3), pp. 343-356

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., Chissom, B. S. (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, *Institute for Simulation and Training, University of Central Florida*.

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I. and Kompatsiaris, Y., 2017. Near-Duplicate Video Retrieval with Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 347-356).

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv preprint* arXiv:1804.08559.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S. and Kompatsiaris, Y., 2017, June. Web Video Verification using Contextual Cues. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security* (pp. 6-10). ACM.

Qazvinian, V., Rosengren, E., Radev, D.R. and Mei, Q., 2011, July. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589-1599). Association for Computational Linguistics.

Teyssou, D, Leung, J.-M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O., and Mezaris, V., 2017, The invid plug-in: web video verification on the browser. In Proceedings of the First Int. Workshop on Multimedia Verification, pp. 23-30. ACM.

Vosoughi, S., Mohsenvand, M.N. and Roy, D., 2017. Rumor gauge: predicting the veracity of rumors on twitter. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *11*(4), p.50.

Wu, K., Yang, S. and Zhu, K.Q., 2015, April. False rumors detection on Sina Weibo by propagation structures. In *31st Int. Conf. on Data Engineering (ICDE),* pp. 651-662. IEEE.

Xie, L., Natsev, A., Kender, J.R., Hill, M. and Smith, J.R., 2011, November. Visual memes in social media: tracking real-world news in youtube videos. In *Proceedings of the 19th ACM international conference on Multimedia* (pp. 53-62). ACM.

Zampoglou, M., Papadopoulos, S. and Kompatsiaris, Y., 2017. Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications*, *76*(4), pp.4801-4834.

Zandi, M., Mahmoudi-Aznaveh, A. and Talebpour, A., 2016. Iterative copy-move forgery detection based on a new interest point detector. *IEEE Transactions on Information Forensics and Security*, *11*(11), pp.2499-2512.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. and Procter, R., 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys (CSUR)*, *51*(2).