Web Video Verification using Contextual Cues

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris
Centre for Research and Technology Hellas, Information Technologies Institute
Thessaloniki, Greece
{olgapapa,markzampoglou,papadop,ikom}@iti.gr

ABSTRACT

As news agencies and the public increasingly rely on User-Generated Content, content verification is vital for news producers and consumers alike. We present a novel approach for verifying Web videos by analyzing their *online context*. It is based on supervised learning on contextual features: one feature set is based on an existing approach for tweet verification adapted to video comments. The other is based on video metadata, such as the video description, likes/dislikes, and uploader information. We evaluate both on a dataset of real and fake videos from YouTube, and demonstrate their effectiveness (F-scores: 0.82, 0.79). We then explore their complementarity and show that under an optimal fusion scheme, the classifier would reach an F-score of 0.9. We finally study the performance of the classifier through time, as more comments accumulate, emulating a real-time verification setting.

CCS CONCEPTS

•**Information systems** → **Information retrieval;**

KEYWORDS

Video verification; context analysis; social media; fake news

ACM Reference format:
Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris. 2017. Web Video Verification using Contextual Cues. In *Proceedings of MFSec'17, Bucharest, Romania, June 06, 2017,* 5 pages.
DOI: http://dx.doi.org/10.1145/3078897.3080535

1 INTRODUCTION AND BACKGROUND

User Generated Content (UGC) currently plays a major role in news reporting. The ubiquitous usage of social media means that non-professionals can contribute vital new information to a news story, including images and videos, that would otherwise be inaccessible to news organizations. However, this new reality also carries risks: UGC originates from unverified sources, and its veracity is by no means guaranteed. Indeed, though the problem of "fake news" gained prominence in the public debate relatively recently, there have already been cases of fake UGC attempting (and often succeeding) to find their way into mainstream news sources in recent years (Figure 1). Thus, there is increasing demand for tools that can assist investigators to timely detect fake content.

MFSec'17, Bucharest, Romania

978-1-4503-5034-1/17/06...$15.00
DOI: http://dx.doi.org/10.1145/3078897.3080535

Figure 1: Thumbnails from recent UGC. Top: two fake videos, "Snake hawk" and "Hezbollah sniper kills ISIS". Bottom: two real videos, "Turkey coup" and "Hetauda rhino".

A significant body of research deals with the automation of news story verification using Web and social media content [9]. One approach is to use tampering detection algorithms for images [8, 10] and videos [5]. Such algorithms have reached a state that allows real-world application with some success [11]. However, in many cases such algorithms fail [10], while in other cases the definition of "fake" extends beyond tampering, thus rendering such algorithms irrelevant. For example, some fake news videos may mislead regarding the context in which they were captured, while others may be staged using studios and actors, and presented as bystander videos. In these cases there is no tampering to detect, and forensics algorithms cannot help. Our study of fake videos has led us to identify five types of *fake* content among news-related UGC:

(1) Staged videos in which actors perform scripted actions under direction, published as UGC.
(2) Videos in which the context of the depicted events is misrepresented (e.g. the claimed video location is wrong).
(3) Past videos presented as UGC from breaking events.
(4) Videos of which the visual or audio content has been altered through editing.
(5) Computer-generated Imagery (CGI) posing as real.

As tampering detection only deals with cases 4 and 5, we need to explore complementary approaches. One such approach is to verify UGC by analyzing the social media posts that disseminate it. There is growing research dedicated to verifying social media posts – most commonly tweets – based on their text, but also from the context around them [4, 7]. In this case, *context* refers to social media activity surrounding the post, such as likes/favorites, shares, replies, as well as the characteristics of the user, such as the number of followers. In the last two years, the MediaEval benchmark has included a "Verifying Multimedia Use" task [1, 3] with the aim of evaluating the veracity of media posted on Twitter using textual, contextual, and image forensics features. The task showed that it

is feasible to assess news-related posts, not only by analyzing the multimedia items or confirming the factual veracity of the posts' claims, but by analyzing the style and context of the posts.

This paper presents a novel approach for detecting fake Web videos by analyzing their context. We adapt a supervised learning approach [2] initially designed to verify Twitter posts using contextual features, to the problem of Web video verification. By relying solely on contextual features instead of visual ones, we can detect fake videos even when tampering detection algorithms fail, in a much faster and more scalable manner. While the principles can be easily applied to most existing video platforms, our implementation is currently based on YouTube due to its popularity. In particular, a *video credibility* descriptor is proposed based on scores produced for YouTube comments using a credibility model originally trained on a corpus of tweets. A second descriptor is based on features directly extracted from YouTube metadata. Experiments demonstrate the effectiveness of our approach. A further contribution includes our analysis of how the verification performance is affected by the evolution of video comments over time.

The main contributions of the work presented in this paper are: a) an annotated dataset of real and fake videos, b) an approach for fake video detection using video comment credibility and video metadata features, and c) an analysis and evaluation of video comment distribution through time, and its effect on video verification.

2 METHOD DESCRIPTION

The method we propose extends our previous work on tweet verification [2] to Web videos, YouTube in particular. We employ a supervised learning approach using two types of feature, and study the potential of fusing them. The first type of feature is based on video comments. Our expectation is that the same characteristics that can help distinguish fake from real tweets can also be applied on YouTube comments. The basis for the proposed approach is a first-level classifier trained on a corpus of tweets annotated as *real* or *fake*. The model is an adaptation from a submission to the MediaEval "Verifying Multimedia Use 2016" Task [2] that proved to be successful in detecting fake tweets. In the original approach, a feature vector is formed from a number of tweet content- and context-based descriptors, and an RBF SVM is trained to classify tweets as "real" or "fake". In order to apply the same protocol to YouTube comments, we need to adapt the features accordingly. For instance, features such as the *number of retweets* are not applicable. Table 1 lists the features used for comment classification. Note that these are only indirectly related to the post veracity. While it would be relatively easy to include telltale features (e.g., the presence of the word "fake" in a tweet or comment), and such a choice could lead to improved success rates in our evaluations, we do not consider that they would be helpful in real-world settings. The reason for this is that we intend to apply our method to videos, of which the veracity cannot be easily assessed by a human, and hence should not rely on users explicitly stating that a video is fake.

In our approach, we first extract these features from a dataset of fake and real tweets, and use them to train a two-class RBF SVM. We then extract the same features from the comments in a dataset of fake and real videos, and use the SVM to classify each comment. This produces a [0,1] value for each comment, corresponding to its credibility score. The estimates for all the comments of a video can then be aggregated to form a video-level descriptor, which can exploit patterns in the comment characteristics, and enable the creation of a second-level model that can classify YouTube videos as real or fake. In our approach, the aggregate descriptor is formed as the 10-bin histogram of first-level estimates of all comments in a video. The [0,1] range is split into 10 bins, and the number of comments with credibility scores falling into each bin is counted. The number of 10 bins was determined experimentally to lead to a low-dimensional but effective representation. The comment credibility histogram then serves as a video-level descriptor. A second-level RBF SVM classifier is trained on these video descriptors to tell fake from real videos.

Table 1: Comment-level features

#	Feature description
01	Text length
02	Number of words
03-04	Contains question/exclamation mark (Boolean)
05-06	Contains happy/sad emoticon (Boolean)
07-09	Contains 1st/2nd/3rd person pronoun (Boolean)
10	Number of uppercase characters
11-12	Number of positive/negative sentiment words
13	Number of slang words
14-15	Has ':' symbol/'please' (Boolean)
16-17	Number of question/exclamation marks
18	Readability score

Table 2: Video-level features

#	Feature description
	From channel description
01	Channel view count
02	Channel comment count
03	Channel subscriber count
04	Channel video count
	From video description
05	Text length
06	Number of words
07-08	Contains question/exclamation mark (Boolean)
09-10	Contains 1st/3rd person pronoun (Boolean)
11	Number of uppercase characters
12-13	Number of positive/negative sentiment words
14	Number of slang words
15	Has ':' symbol (Boolean)
16-17	Number of question/exclamation marks

The other type of feature is extracted directly from the video metadata. These include features that describe the uploader channel, and also text-based features from the video description. These are then used to train a classifier to distinguish fake from real videos directly based on their context. Table 2 presents the 17 features we extract from the video metadata. They are inspired by the ones in Table 1, but we have kept only those that can be applied to the video channel information and the video metadata. Thus, the video descriptor consists of 17 boolean and integer values, which are then used to train an RBF SVM classifier. In selecting both the comment and the video descriptors, we ignored features that took uniform values for all items in the dataset. For example, the feature "Contains 2nd person pronoun" was *false* for all video descriptions, and thus was not included in the list of video features.

3 EXPERIMENTS

3.1 Datasets

The comment-level classifier requires a large dataset of fake and real tweets or comments for training. The Image Verification Corpus[1] (IVC) is such a dataset. It contains 17,857 tweets from 53 past events, 7,229 of which have been labeled real and 10,628 fake, and was used for the MediaEval Verifying Multimedia Use tasks.

Video-level classification requires an annotated dataset of fake and real videos. To this end, we created the Fake Video Corpus (FVC), a collection of UGC YouTube videos from the recent past around news stories that have been verified as factual. The set contains 55 *fake* videos and 49 *real* ones. Each video contains on average 856 comments. A first version of the FVC – containing the fake videos – is already publicly available [6], and we intend to upload an updated version which will include the real ones. The example videos shown in Figure 1 are indicative of the dataset.

3.2 Evaluation

Concerning the comment-based features, the first set of evaluations dealt with their ability to distinguish between real and fake UGC videos in their current state, i.e. using all the comments that have been accumulated from the first posting until "now". The comment-level classifier was trained on the IVC, and then used to produce credibility scores for all comments of the FVC videos. For each video, the scores of its comments were aggregated into a 10-bin histogram serving as a descriptor for the entire video. The FVC dataset was split into training and test sets using 10-fold cross-validation to evaluate the performance of the comment-based approach. Results are shown in Table 3 ("Comments" row).

With respect to the video metadata approach, the features of Table 2 were extracted from all videos of the FVC, and an RBF SVM classifier was trained and evaluated on the dataset using 10-fold cross-validation. Results are shown in the "Video metadata" row of Table 3. The third row also shows the results of an ideal fusion method having an oracle select the correct classifier: this will make a mistake only when both classifiers make a mistake. The oracle-based fusion yields a significant increase in performance, which means that the two classifiers are complementary, and we would benefit by combining them under an appropriate scheme. However, as our initial attempts at combining the two models using simple feature concatenation (early fusion) did not yield significant success as yet, how to perform fusion effectively is still an open question.

Table 3: Video classification results

	Precision	Recall	F1
Comments	0.88	0.74	0.79
Video metadata	0.88	0.79	0.82
Ideal fusion	1.00	0.83	0.90

Despite the encouraging results from these first tests, there is a caveat: these videos were found to be fake quite a long time ago, and this will most likely have affected the characteristics of their comments, as many users have written their comments after they already knew that the video is fake. For the needs of breaking news reporting, this is not a realistic usage scenario, as generally we need to be able to evaluate a video during the first minutes or at most hours after it is posted.

[1] https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2016

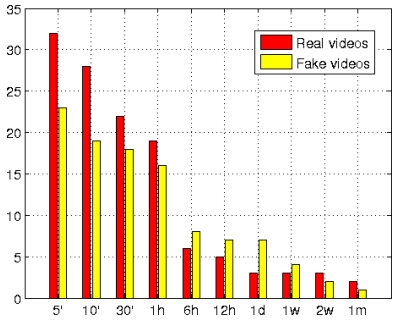

Figure 2: Number of videos with zero comments after specific time intervals.

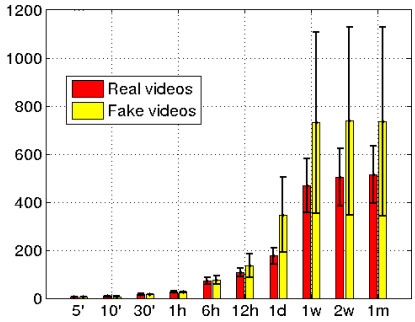

Figure 3: Average number of comments per video through time. Standard deviation bars were scaled down by 10 for the sake of clarity.

As a result, we ran a second set of evaluations based on this principle. We defined a number of time limits (e.g., 5', 30', hour, day, week) after the video was first posted, and only retained for each video the comments that were made before each limit. Out of the 55 fake and 49 real videos, Figure 2 shows the number of videos having zero comments (which were left out of the evaluation). This is one limitation of our approach, as the comment-based classifier cannot operate on videos without comments. Figure 3 shows the average number of comments per video. It should be noted that the standard deviation bars are scaled down by 10, which indicates an extremely high variance in the number of comments. For the evaluation at different time frames, we tested two approaches. One is to train a classifier using all available comments in the training set (*single classifier*), and use this model to classify videos at any time frame. The other is to train one classifier for each time frame (*multiple classifiers*), using those comments in the training set that correspond to that time, under the assumption that the passing of time does not affect only the number of comments but also the distribution of their credibility values. Figure 4 shows the results for the two approaches. It can be seen that the single classifier performs better early on, although still quite low (F1-score<0.5). At 6 hours the multiple classifiers approach performs better, reaching F1=0.73, and the two approaches ultimately converge.

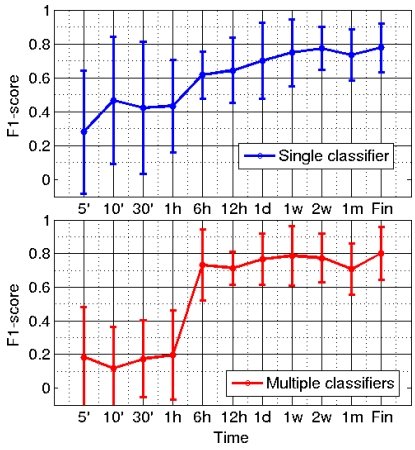

Figure 4: Comparison between the single and multiple classifiers approaches.

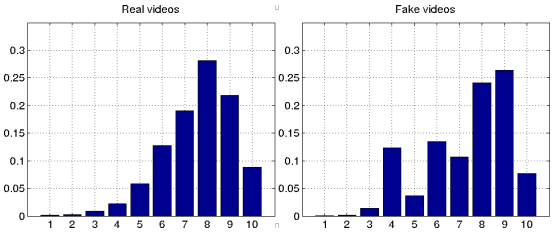

Figure 5: Comparison of comment credibility estimate distributions between real and fake videos.

In trying to interpret the classification results, we attempted to study the feature value distribution, and see if any significant differences appear between the two classes. Figure 5 shows the mean comment histogram feature for all "fake" and "real" videos. By comparing the histograms at different times, we observed that the overall distribution remains fundamentally unchanged as the comment sections evolve, thus we are only presenting the results for all comments (i.e. after convergence). It is clear that the two sets exhibit differences. Histogram values for real videos seem to follow a unimodal distribution, peaking at bin 8, while for fake videos there is a peak at 9, and two smaller ones at bins 4 and 6. It is currently difficult to interpret the cause of these differences. It appears that a classifier trained on tweets tends to return higher values (centered at 0.8-0.9) for YouTube comments, corresponding to a high probability of the comment being assigned to the "fake" class. However, it is also clear that a significant percentage of low-value (high credibility) comments appear under fake videos - these could actually be user comments debunking the video, but a deeper analysis would be necessary to confirm this. The code and data necessary to reproduce our experiments can be found on GitHub[2].

4 CONCLUSIONS

We presented a methodology for automatically assessing the veracity of YouTube videos based on supervised classification using two types of feature. One builds upon well established features for detecting fake Twitter posts, which we adapted to YouTube comments, while the other is based on similar features extracted from the metadata of a YouTube video and its channel. While the approach was tested exclusively on YouTube videos, it can be extended to other platforms providing similar context (video descriptions, user profile, comments) such as Dailymotion or Facebook. It is in our future plans to adapt and test the approach to such platforms. Evaluations demonstrated the effectiveness of both features in telling apart fake from real videos, without having to analyze the actual visual content. Furthermore, we demonstrated the complementarity of the two features, which suggests that, under a reliable fusion scheme, we could achieve even better results.

However, the small size of our dataset and the overall difficulty of collecting real-world, relevant videos for both categories – *real* and *fake* – does not currently allow us to perform more exhaustive evaluations, or train an optimal fusion model. Further evaluations on the time needed to accumulate enough comments for successful detection demonstrated that, while during the first hour we do not usually have enough information to produce a reliable estimate, performance picks up relatively quickly, and can lead to relatively reliable estimates during the first six hours after the posting. As the video description and metadata already provide reliable results from the moment the video is posted, it seems more reasonable to rely more on such features early on, while our current comment features appear to become useful at a later stage.

Given the encouraging initial results of our approach, our plan is to extend it with the aim of offering a tool for verification assistance, targeted at news professionals and laypeople alike. One step in this direction would be to extend the Fake Video Corpus, to allow for a wider training set and more exhaustive evaluations. Manually identifying fake and real videos is a labor-intensive task. However, both the comment- and video-based models could benefit from a larger training set. Furthermore, a larger set would allow more reliable evaluations leading to deeper insights on the strengths and weaknesses of the approaches.

Another step would be to identify which features really contribute to the classification. Our preliminary analysis showed that there exist significant differences in the comment credibility estimate distributions between the two sets. However, a targeted analysis is necessary to identify how the tweet classification model translates to YouTube comments.

Furthermore, we intend to work towards a fusion scheme between the two features, to exploit their complementarity towards increasing the overall accuracy. We believe this is partly dependent on broadening the current video dataset, to allow for more elaborate fusion schemes, since our initial efforts towards using a simple approach such as feature concatenation have not yielded any significant improvement. Building a large enough set to train a specialized fusion model seems a promising approach, given the performance of the oracle scheme.

5 ACKNOWLEDGMENTS

This work is supported by the InVID project, which is funded by the European Commission under contract number 687786.

[2]https://github.com/MKLab-ITI/contextual-video-verification/tree/master/MFSec_2017

REFERENCES

[1] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2015. Verifying Multimedia Use at MediaEval 2015. In *MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany*.

[2] Christina Boididou, Stuart E. Middleton, Symeon Papadopoulos, Dang Nguyen, Duc Tien, Michael Riegler, Giulia Boato, Andreas Petlund, and Yiannis Kompatsiaris. 2016. The VMU Participation @ verifying multimedia use 2016. In *MediaEval Benchmarking Initiative for Multimedia Evaluation 2016*. CEUR-WS.

[3] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E. Middleton, Andreas Petlund, and Yiannis Kompatsiaris. 2016. Verifying Multimedia Use at MediaEval 2016. In *MediaEval*, Vol. 1739. CEUR-WS.org.

[4] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *22nd International World Wide Web Conference, WWW '13*. ACM, 729–736.

[5] Ramesh Chand Pandey, Sanjay Kumar Singh, and Kaushal K. Shukla. 2016. Passive forensics in image and video using noise features: A review. *Digital Investigation* 19 (2016), 1–28. http://dx.doi.org/10.1016/j.diin.2016.08.002

[6] Symeon Papadopoulos, Markos Zampoglou, Ioannis Kompatsiaris, and Denis Teyssou. 2017. InVID Fake Video Corpus. (Jan. 2017). DOI:https://doi.org/10.5281/zenodo.242481

[7] Meet Rajdev and Kyumin Le. 2015. Fake and Spam Messages: Detecting Misinformation During Natural Disasters on Social Media. In *WI-IAT (1)*. IEEE Computer Society, 17–20. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7395995

[8] Ewerton Silva, Tiago Jose de Carvalho, Anselmo Ferreira, and Anderson Rocha. 2015. Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes. *J. Visual Communication and Image Representation* 29 (2015), 16–32. http://dx.doi.org/10.1016/j.jvcir.2015.01.016

[9] Neil Thurman, Steve Schifferes, Richard Fletcher, Nic Newman, Stephen Hunt, and Aljosha Karim Schapals. 2016. Giving Computers a Nose for News. *Digital Journalism* 4, 7 (2016), 838–848.

[10] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. A Large-Scale Evaluation of Splicing Localization Algorithms for Web Images. *Multimedia Tools and Applications* 76, 4 (February 2017), 4801––4834.

[11] Markos Zampoglou, Symeon Papadopoulos, Yiannis Kompatsiaris, Ruben Bouwmeester, and Jochen Spangenberg. 2016. Web and Social Media Image Forensics for News Professionals. In *SMN@ICWSM (AAAI Workshops)*, Jisun An, Haewoon Kwak, and Fabrício Benevenuto (Eds.), Vol. WS-16-19. AAAI Press. http://www.aaai.org/Library/Workshops/ws16-19.php