Verifying Multimedia Use at MediaEval 2016

Christina Boididou[1], Symeon Papadopoulos[1], Duc-Tien Dang-Nguyen[2], Giulia Boato[2], Michael Riegler[3], Stuart E. Middleton[4], Andreas Petlund[3], and Yiannis Kompatsiaris[1]

[1]Information Technologies Institute, CERTH, Greece. [boididou,papadop,ikom]@iti.gr

[2]University of Trento, Italy. [dangnguyen,boato]@disi.unitn.it

[3]Simula Research Laboratory, Norway. michael@simula.no,apetlund@ifi.uio.no

[4]University of Southampton IT Innovation Centre, Southampton, UK. sem@it-innovation.soton.ac.uk

ABSTRACT

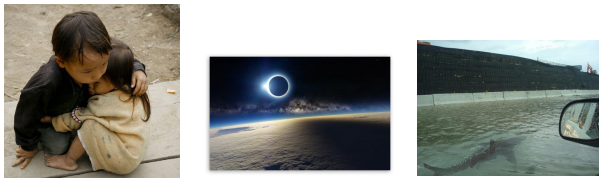
This paper provides an overview of the Verifying Multimedia Use task that takes places as part of the 2016 MediaEval Benchmark. The task motivates the development of automated techniques for detecting manipulated and misleading use of web multimedia content. Splicing, tampering and reposting videos and images are examples of manipulation that are part of the task definition. For the 2016 edition of the task, a corpus of images/videos and their associated posts is made available, together with labels indicating the appearance of misuse (`fake`) or not (`real`) in each case as well as some useful post metadata.

1. INTRODUCTION

Social media, such as Twitter and Facebook, as means of news sharing is very popular and also very often used by government or politicians to reach the public. The speed of news spreading on such platforms often leads to the appearance of large amounts of misleading multimedia content. Given the need for automated real-time verification of this content, several techniques have been presented by researchers. For instance, previous work focused on the classification between fake and real tweets spread during Hurricane Sandy [6] and other events [2] or on automatic methods for assessing posts' credibility [3]. Several systems for checking content credibility have been proposed, such as Truthy [8], TweetCred [5] and Hoaxy [9]. The second edition of this task aims to encourage the development of new verification approaches. This year, the task is extended by introducing a sub-task, focused on identifying digitally manipulated multimedia content. To this end, we encourage participants to create text-focused and/or image-focused approaches equally.

2. TASK OVERVIEW

Main task. The definition of the main task is the following: "Given a social media post, comprising a text component, an associated piece of visual content (image/video) and a set of metadata originating from the social media platform, the task requires participants to return a decision (fake, real or unknown) on whether the information presented by this post sufficiently reflects the reality." In practice, participants receive a list of posts that are associated with images and are required to automatically predict, for each post, whether it is trustworthy or deceptive (`real` or `fake` respectively). In addition to fully automated approaches, the task also considers human-assisted approaches provided that they are practical (i.e., fast enough) in real-world settings. The following definitions should be also taken into account:



(a) (b) (c)

Figure 1: Examples of misleading (fake) image use: (a) reposting of real photo claiming to show two Vietnamese siblings at Nepal 2015 earthquake; (b) reposting of artwork as a photo of solar eclipse (March 2015); (c) spliced sharks on a photo during Hurricane Sandy in 2012.

- A post is considered **fake** when it shares multimedia content that does not represent the event that it refers to. Figure 1 presents examples of such content.
- A post is considered **real** when it shares multimedia that legitimately represents the event it refers to.
- A post that shares multimedia content that does not represent the event it refers to but reports the false information or refers to it with a sense of humour **is neither considered fake nor real** (and hence not included in the task dataset).

Sub-task. This version of the task addresses the problem of detecting digitally manipulated (tampered) images. The definition of the task is the following: "Given an image, the task requires participants to return a decision (`tampered`, `non-tampered` or `unknown`) on whether the image has been digitally modified or not". In practice, participants receive a list of images and are required to predict if this image is tampered or not, using multimedia forensic analysis. It should also be noted that an image is considered **tampered** when it is digitally altered.

In both cases, the task also asks participants to optionally return an explanation (which can be a text string, or URLs pointing to evidence) that supports the verification decision.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands

The explanation is not used for quantitative evaluation, but rather for gaining qualitative insights into the results.

3. VERIFICATION CORPUS

Development dataset (`devset`): This is provided together with ground truth and is used by participants to develop their approach. For the *main task*, it contains posts related to the 17 events of Table 1, comprising in total 193 cases of real and 220 cases of misused images/videos, associated with 6,225 real and 9,596 fake posts posted by 5,895 and 9,216 unique users respectively. This data is the union of last year's `devset` and `testset` [1]. Note that several of the events, e.g., Columbian Chemicals and Passport Hoax are hoaxes, hence all multimedia content associated with them is misused. For several real events (e.g., MA flight 370) no real images (and hence no `real` posts) are included in the dataset, since none came up as a result of the data collection process that is described below. For the *sub-task*, the development set contains 33 cases of non-tampered and 33 cases of tampered images, derived from the same events, along with their labels (`tampered` and `non-tampered`).

Test dataset (`testset`): This is used for evaluation. For the *main task*, it comprises 104 cases of real and misused images and 25 cases of real and misused videos, in total associated with 1,107 and 1,121 posts, respectively. For the *sub-task*, it includes 64 cases of both tampered and non-tampered images from the testset events.

The data for both datasets are publicly available[1]. Similar to the 2015 edition of the task, the posts were collected around a number of known events or news stories and contain fake and real multimedia content manually verified by cross-checking online sources (articles and blogs). Having defined a set of keywords K for each `testset` event, we collected a set of posts P (using Twitter API and specific keywords) and a set of unique fake and real pictures around these events, resulting in the fake and real image sets I_F, I_R respectively. We then used the image sets as seeds to create our reference verification corpus $P_C \subset P$, which includes only those posts that contain at least one image of the predefined sets I_F, I_R. However, in order not to restrict the posts to the ones pointing to the exact image, we employed a scalable visual near-duplicate search strategy [10]: we used the I_F, I_R as visual queries and for each query we checked whether each post image from the P set exists as an image item or a near-duplicate image item of the I_F or the I_R set. In addition to this process, we also used a real-time system that collects posts using keywords and a location filter [7]. This was performed mainly to increase the real samples for events that occurred in known locations.

To further extend the `testset`, we carried out a crowdsourcing campaign using the microWorkers platform[2]. We asked each worker to provide three cases of manipulated multimedia content that they found on the web. Furthermore, they had to provide a link with information and description on each case, along with online resources containing evidence of its misleading nature. We also asked them to provide the original content if available. To avoid cheating, they had to provide a manual description of the manipulation. We also tested the task in two pilot studies to be sure that the information we got would also be useful. Overall, the data collected was very useful. We performed 75 tasks and each worker earned $2,75$\$ per task.

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

[2] `https://microworkers.com/`

Table 1: `devset` events: For each event, we report the numbers of unique real (if available) and fake images/videos (I_R, I_F respectively), unique posts that shared those images (P_R, P_F) and unique Twitter accounts that posted those tweets (U_R, U_F).

Name	I_R	P_R	U_R	I_F	P_F	U_F
Hurricane Sandy	148	4,664	4,446	62	5,559	5,432
Boston Marathon bombing	28	344	310	35	189	187
Sochi Olympics	-	-	-	26	274	252
Bring Back Our Girls	-	-	-	7	131	126
MA flight 370	-	-	-	29	501	493
Columbian Chemicals	-	-	-	15	185	87
Passport hoax	-	-	-	2	44	44
Rock Elephant	-	-	-	1	13	13
Underwater bedroom	-	-	-	3	113	112
Livr mobile app	-	-	-	4	9	9
Pig fish	-	-	-	1	14	14
Nepal earthquake	11	1004	934	21	356	343
Solar Eclipse	4	140	133	6	137	135
Garissa Attack	2	73	72	2	6	6
Samurai and Girl	-	-	-	4	218	212
Syrian Boy	-	-	-	1	1786	1692
Varoufakis and ZDF	-	-	-	1	61	59
Total	193	6225	5895	220	9596	9216

For every item of the datasets, we extracted and made available three types of features, similar to the ones we made available for the 2015 edition of the task: (i) **features extracted from the post itself**, i.e., the number of words, hashtags, mentions, etc. in the post's text [1], (ii) **features extracted from the user account**, i.e., number of friends and followers, whether the user is verified, etc. [1]. and (iii) **forensic features extracted from the image**, i.e., the probability map of the aligned double JPEG compression, the estimated quantization steps for the first six DCT coefficients of the non-aligned JPEG compression, and the Photo-Response Non-Uniformity (PRNU) [4].

4. EVALUATION

Overall, the *main task* is interested in the accuracy with which an automatic method can distinguish between use of multimedia in posts in ways that faithfully reflect reality versus ways that spread false impressions. Hence, given a set of labelled instances (post + image + label) and a set of predicted labels (included in the submitted runs) for these instances, the classic IR measures (i.e., Precision P, Recall R, and F-score) are used to quantify the classification performance, where the target class is the class of `fake` tweets. Since the two classes (`fake`/`real`) are represented in a relatively balanced way in the `testset`, the classic IR measures are good proxies of the classifier accuracy. Note that task participants are allowed to classify a tweet as `unknown`. Obviously, in case a system produces many `unknown` outputs, it is likely that its precision will benefit, assuming that the selection of `unknown` is done wisely, i.e. successfully avoiding erroneous classifications. However, the recall of such a system will suffer in case the tweets that are labelled as `unknown` turn out to be `fake` (the target class). Similarly, in the *sub-task* case, given the instances of (image + label), we use the same IR measures to quantify the performance of the approach, where the target class is `tampered`.

5. ACKNOWLEDGEMENTS

This work is supported by the REVEAL and InVID projects, partially funded by the European Commission (FP7-610928 and H2020-687786 respectively).

6. REFERENCES

[1] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris. Verifying multimedia use at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[2] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 743–748. ACM, 2014.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[4] V. Conotter, D.-T. Dang-Nguyen, M. Riegler, G. Boato, and M. Larson. A crowdsourced data set of edited images online. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, CrowdMM '14, pages 49–52, New York, NY, USA, 2014. ACM.

[5] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[6] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking Sandy: characterizing and identifying fake images on twitter during Hurricane Sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 729–736, 2013.

[7] S. E. Middleton, L. Middleton, and S. Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17, 2014.

[8] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[9] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 745–750. International World Wide Web Conferences Steering Committee, 2016.

[10] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over VLAD and Product Quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 16(6):1713–1728, 2014.