Verifying Multimedia Use at MediaEval 2015

Christina Boididou[1], Katerina Andreadou[1], Symeon Papadopoulos[1], Duc-Tien Dang-Nguyen[2], Giulia Boato[2], Michael Riegler[3], and Yiannis Kompatsiaris[1]

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

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

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

ABSTRACT

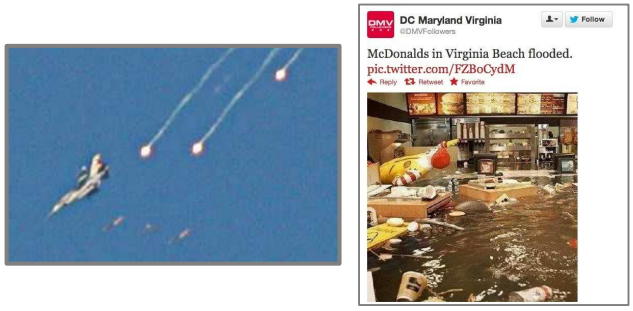
This paper provides an overview of the Verifying Multimedia Use task that takes places as part of the 2015 MediaEval Benchmark. The task deals with the automatic detection of manipulation and misuse of Web multimedia content. Its aim is to lay the basis for a future generation of tools that could assist media professionals in the process of verification. Examples of manipulation include maliciously tampering with images and videos, e.g., splicing, removal/addition of elements, while other kinds of misuse include the reposting of previously captured multimedia content in a different context (e.g., a new event) claiming that it was captured there. For the 2015 edition of the task, we have generated and made available a large corpus of real-world cases of images that were distributed through tweets, along with manually assigned labels regarding their use, i.e. misleading (`fake`) versus appropriate (`real`).

1. INTRODUCTION

Modern Online Social Networks (OSN), such as Twitter, Instagram and Facebook, are nowadays the primary sources of information and news for millions of users and the major means of publishing user-generated content. With the growing number of people participating and contributing to these communities, analyzing and verifying the massive amounts of such content has emerged as a major challenge. Veracity is a crucial aspect of media content, especially in cases of breaking news stories and incidents related to public safety, ranging from natural disasters and plane crashes to terrorist attacks. Popular stories have such profound impact on the public attention that content gets immediately retransmitted by millions of users, and often it is found to be misleading, resulting in misinformation of the public audience and even of the authorities.

In this setting, there is increasing need for automated real-time verification and cross-checking tools. Work has been done in this field and techniques for evaluating tweets have been proposed. Gupta et al. [4] used the *Hurricane Sandy* natural disaster case to highlight the role of Twitter in spreading fake content during the event, and proposed classification models to distinguish between fake and real tweets. Ito et al. [5] proposed a method to assess tweet credibility by using "tweet-" and "user"-topic features derived from the Latent Dirichlet Allocation (LDA) model. They also introduce the user's "expertness" and "bias" features, demonstrating that the bias features work better. Given the importance of the problem, as attested by the increasing number of works in the area [3], this task aspires to foster the development of new Web multimedia verification approaches.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany



(a) Manipulation (b) Reposting

Figure 1: Examples of `fake` web multimedia: a) digitally manipulated image of an IAF F-16 deploying a flare over Southern Lebanon; the flare was digitally duplicated; b) an image posted during Hurricane Sandy that is a repost from a 2009 art installation.

2. TASK OVERVIEW

The definition of the task is the following: "Given a tweet and the accompanying multimedia item (image or video) from an event that has the profile to be of interest in the international news, return a binary decision representing verification of whether the multimedia item reflects the reality of the event in the way purported by the tweet." In practice, participants received a list of tweets that include images and were required to automatically predict, for each tweet, whether it is trustworthy or deceptive (`real` or `fake` respectively). In addition to fully automated approaches, the task also considered human-assisted approaches provided that they are practical (i.e., fast enough) in real-world settings. The following considerations should be made in addition to the above definition:

- A tweet 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 tweet is considered **real** when it shares multimedia that legitimately represents the event it refers to.
- A tweet 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 datasets released by the task).

The task also asked participants to optionally return an explanation (which can be a text string, or URLs pointing to resources online) that supports the verification decision. The explanation was not used for quantitative evaluation, but rather for gaining qualitative insights into the results.

3. VERIFICATION CORPUS

Development dataset (`devset`): This was provided together with ground truth and used by participants to develop their approach. It contains tweets related to the 11 events of Table 1, comprising in total 176 cases of real and 185 cases of misused images, associated with 5,008 real and 7,032 fake tweets posted by 4,756 and 6,769 unique users respectively. Note that several of the events, e.g., Columbian Chemicals, Passport Hoax and Rock Elephant, were actually hoaxes, hence all multimedia content associated with them was misused. For several real events (e.g., MA flight 370) no real images (and hence no `real` tweets) were included in the dataset, since none came up as a result of the data collection process that is described below.

Test dataset (`testset`): This was used for evaluation. It comprises 17 cases of real images, 33 of misused images and 2 cases of misused videos, in total associated with 1,217 real and 2,564 fake tweets that were posted by 1,139 and 2,447 unique users respectively.

The tweet IDs and image URLs for both datasets are publicly available[1]. Both consist of tweets collected around a number of widely known events or news stories. The tweets contain fake and real multimedia content that has been manually verified by cross-checking online sources (articles and blogs). The data were retrieved with the help of Topsy and Twitter APIs using keywords and hashtags around these specific events. Having defined a set of keywords K for each event of Table 1, we collected a set of tweets T. Afterwards, with the help of online resources, we identified a set of unique fake and real pictures around these events, and created the fake and the real image sets I_F, I_R respectively. We then used the image sets as seeds to create our reference verification corpus $T_C \subset T$. This corpus includes only those tweets that contain at least one image of the predefined sets of images I_F, I_R. However, in order not to restrict the tweets to only those that point to the exact seed image URLs, we also employed a scalable visual near-duplicate search strategy as described in [6]. More specifically, we used the sets of fake and real images as visual queries and for each query we checked whether each image tweet from the T set exists as an image item or a near-duplicate image item of the I_F or the I_R set. To ensure near-duplicity, we empirically set a minimum threshold of similarity tuned for high precision. However, a small amount of the images exceeding the threshold turned out to be irrelevant to the ones in the seed set. To remove those, we conducted a manual verification step on the extended set of images.

[1] `https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015`

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

Name	I_R	T_R	U_R	I_F	T_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
MA flight 370	-	-	-	29	501	493
Bring Back Our Girls	-	-	-	7	131	126
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
Total	176	5,008	4,756	185	7,032	6,769

For every item of the aforementioned datasets, we extracted and made available three types of features:

- Features extracted from the tweet itself, for instance the number of terms, the number of URLs, hashtags, the number of mentions, etc. [1].
- User-based features which are based on the Twitter user profile, for instance the number of friends and followers, the number of times the user is included in a Twitter list, whether the user is verified, etc. [1].
- Forensic features extracted from the visual content of the tweet image, for instance the probability map of the aligned double JPEG compression, the potential primary quantization steps for the first six DCT coefficients of the non-aligned JPEG compression, and the PRNU (Photo-Response Non-Uniformity) [2].

4. EVALUATION

Overall, the task is interested in the accuracy with which an automatic method can distinguish between use of multimedia in tweets in ways that faithfully reflect reality versus ways that spread false impressions. Hence, given a set of labelled instances (tweet + 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) were 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 were 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` was done wisely, i.e. successfully avoiding erroneous classifications. However, the recall of such a system would suffer in case the tweets that were labelled as `unknown` turned out to be `fake` (the target class).

5. ACKNOWLEDGEMENTS

We would like to thank Martha Larson for her valuable feedback in shaping the task and writing the overview paper. This work is supported by the REVEAL project, partially funded by the European Commission (FP7-610928).

6. REFERENCES

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

[2] 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.

[3] N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2451–2460, New York, NY, USA, 2012. ACM.

[4] 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.

[5] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama. Assessment of tweet credibility with LDA features. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 953–958, 2015.

[6] 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.