A HYBRID APPROACH FOR MULTIMEDIA USE VERIFICATION

Quoc-Tin Phan[1], Alessandro Budroni[2], Cecilia Pasquini[1], Francesco G. B. De Natale[3]
Department of Information Engineering and Computer Science - University of Trento, Italy
{quoctin.phan, cecilia.pasquini}@unitn.it[1]; alessandro.budroni@studenti.unitn.it[2]; denatale@ing.unitn.it[3]

ABSTRACT

Social networks enable multimedia sharing between worldwide users, however, there is no automatic mechanism implemented aiming to verifying multimedia use. This has been known as a highly challenging problem due to the variety of media types and huge amount of information they convey. As a participating team of MediaEval 2016, we propose a hybrid approach for detecting misused multimedia on Twitter which has been known as Verifying Multimedia Use task. Specifically, we designed a verfication system that can answer how likely an associated multimedia file is fake based on multiple forensic features and textual features, which were acquired by performing online text search and image reverse search. Next, effective post-based features and user-based features are utilized to validate the credibility of tweet posts. Finally, based on the assumption that a tweet sharing fake images or videos is likely to be fake, credibility scores of tweet posts and associated multimedia are fused to detect misused multimedia.

1. INTRODUCTION

Online Social Network (OSN) services offer a medium for users to connect and share daily information. With respect to specific events, part of information is usually not trustable and its dissemination causes several negative consequences on the community. Attempts have been proposed to address the problem of image manipulation on online news [9], or the impact of image manipulations to users' perceptions [6].

In MediaEval Verifying Multimedia Use task [3, 5] , given tweet content features, user features and some effective forensic features, innovative methods are welcomed to verify whether multimedia (images and videos) are correctly used on Twitter. Due to the variety of languages used and the fact that many reposted tweets do not contain meaningful textual information, linguistic approaches like [8, 10] are believed not effective enough in this task. Moreover, almost each tweet post is accompanied by at least an image or video, and the image or video itself reflects the credibility of tweet. To the best of our knowledge, only [4] took into account multimedia forensic features in Multimedia Use Verification task.

Despite the fact that associated multimedia files play a significant role in assessing credibility of tweets, forensic algorithms are very sensitive to subsequent image modifications and multiple lossy compression. In this work we propose a novel approach to assess the credibility of associated images or videos by using not only forensic features but also textual features which are acquired by performing online text search and image reverse search. The acquired results on development and test sets confirm the effectiveness of our proposed method.

2. THE PROPOSED METHOD

We propose a verification system composing two classification tiers as depicted in Figure 1. The first classification tier takes as inputs the event and the associated image or video, and answer *How likely does this image or video reflect the event?*. We consider the occurence context of associated images or videos on the Internet as a strong evidence for assessing their trustworthiness. Having certain confidence about the credibility of associated images or videos, we proceed to design the second classification tier to validate the credibility of tweets based on Twitter-based features. Finally, scores returned from two classifiers are fused to give final decision.

2.1 Multimedia assessment

In the first step, we conduct online text search using relevant keywords associated with the event and select top returned websites from which we extract most relevant terms based on the statistical measurement TF-IDF (Term Frequency - Inverse Document Frequency). On another side, the associated image is searched over Google Images and we select only top returned websites to check the frequency of most relevant terms from event text search. To Youtube videos, only users' comments are extracted, while leaving out videos from other cites unprocessed. By this step, the system is expected to correctly recognize images or videos not belonging to current event. In the second step, we check occurences of positive, negative and "fake" related words in the whole text retrieved from image or video reverse search, assuming that a fake multimedia should receive negative assessment from readers.

Forensic operations can be applied on multimedia files to verify whether or not the multimedia file is tampered, and even which regions are most likely to be modified. We adopt non-aligned double JPEG compression [2], block artifact grid [7], and Error Level Analysis [1] as useful forensic features. Finally, we integrate textual features and forensic features in the first classification tier.

2.2 Tweet credibility assessment

After having the output from the first classification tier

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

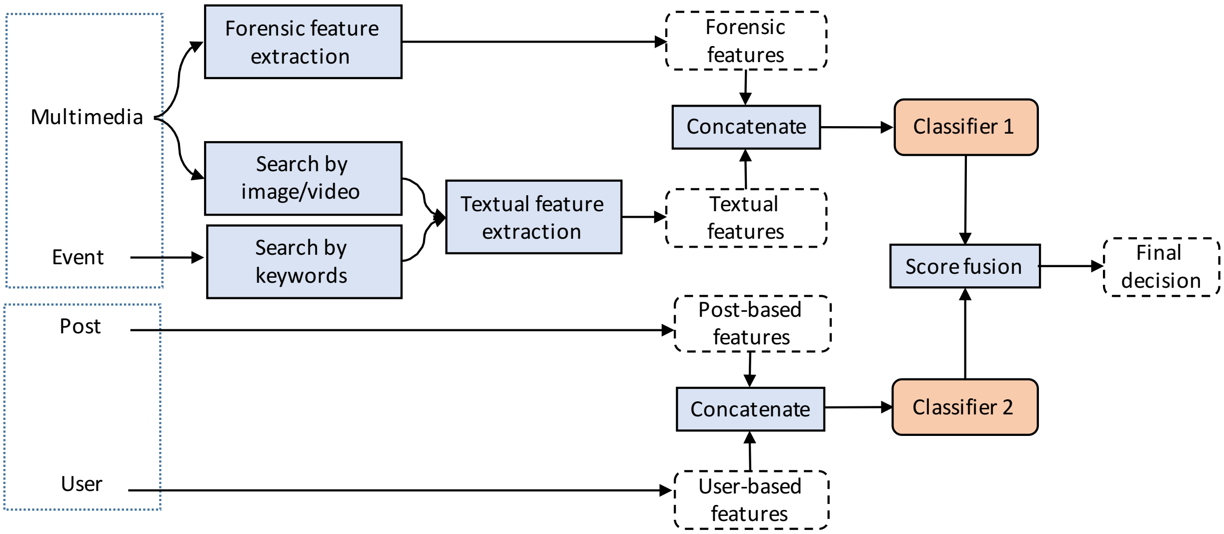

Figure 1: Schema of the proposed method.

reflecting the trustworthiness of associated multimedia, the second classification tier is designed to assess how multimedia are used on Twitter. Tweet credibility assessment is feasible thanks to post-based features, i.e. whether the tweet contains the question mark or exclamation mark characters, number of negative sentiment words the tweet contains, together with user-based features, i.e. the number of followers the user has, whether the user is verified by Twitter.

2.3 Score fusion

We approach the problem by experimenting with LR (Logistic Regression) and RF (Random Forest) classifiers. As depicted in Table 1, LR performs less efficient than RF on the development set. This can be explained as RF suits well with non-linearly separable and uneven data, i.e. some Twitter posts do not associate with any meaningful text, forensic features of videos are not included (all are zeros). For that reason, we select RF as our classifiers and proceed to final decision by conducting score level fusion. With the assumption that *a tweet sharing fake images or videos is likely to be fake*, higher weight is assigned to the output of the first tier, while lower weight to the second tier. In order to validate our method, we conduct experiments counting only scores from classification tier 2 (using post-based and user-based features provided by the task), and experiments using 0.8 : 0.2 weighting strategy. Statistics shown in Table 1 confirm the effectiveness of our multimedia assessment tier and score fusion strategy.

Table 1: Verification results on the development set in terms of F1-score, 100 real and 100 fake samples selected from {*Hurricane Sandy, Boston Marathon Blast, Nepal Earthquake*} for training, 300 real and 300 fake samples from other events for testing.

	LR	RF
Tier 2 scores	0.44	0.54
Fused scores	0.81	**0.88**

3. RESULTS AND DISCUSSION

In this section, we report accumulated results on the subtask based on our multimedia assessment approach and the main task based on two-tier classification. In the sub-task, we submit two RUNs: i) RUN 1 (required): apply only forensic features described in Section 2.1, ii) RUN 2: apply both textual features and forensic features described in Section 2.1. Especially, on the second RUN, we train the classifier on entire multimedia available in development set of the main task. Acquired results from Table 2 reveal the fact that our method gains *recall* if we take into account textual features acquired from online text search and image reverse search. This means we can effectively reduce false negative rate and more fake samples are detected.

Table 2: Verification results on the test set of the sub-task

	Recall	Precision	F1-score
RUN 1	0.5	0.48	0.49
RUN 2	0.93	0.49	**0.64**

Next, results of the main task are reported from three RUNs: i) RUN 1 (required): apply only the second classification tier, ii) RUN 2: apply two-tier classification and 0.8 : 0.2 fusion strategy, answer *UNKNOWN* to cases where the output of classification tier 1 is not available due to online searching errors, iii) RUN 3: apply two-tier classification and 0.8 : 0.2 fusion strategy, consider only the output of classification 2 to cases where the output of classification tier 1 is not available due to online searching errors.

Table 3: Verification results on the test set of the main task

	Recall	Precision	F1-score
RUN 1	0.55	0.71	0.62
RUN 2	0.94	0.81	**0.87**
RUN 3	0.94	0.74	0.83

Results from Table 3, especially RUN 2, again confirms the effectiveness of our proposed method on multimedia assessment and fusion strategy. Our method, however, is subject to online searching errors which happen to videos NOT hosted by YouTube.

4. REFERENCES

[1] Error level analysis tutorial. http://fotoforensics.com/tutorial-ela.php. Accessed: 28/08/2016.

[2] T. Bianchi and A. Piva. Image Forgery Localization via Block-Grained Analysis of JPEG Artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3):1003–1017, June 2012.

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

[4] C. Boididou, S. Papadopoulos, D.-T Dang-Nguyen, G. Boato, and Y. Kompatsiaris. The CERTH-UNITN Participation @ Verifying Multimedia Use 2015. In *Proceedings of the MediaEval 2015 Workshop*, pages 6–8, 2015.

[5] C. Boididou, S. Papadopoulos, D.-T Dang-Nguyen, G. Boato, M. Riegler, S. E. Middleton, A. Petlund, and Y. Kompatsiaris. Verifying Multimedia Use at MediaEval 2016. In *Proc. of the MediaEval 2016 Workshop*, Hilversum, Netherlands, Oct. 20-21 2016.

[6] V. Conotter, D.-T. Dang-Nguyen, G. Boato, M. Menéndez, and M. Larson. Assessing the impact of image manipulation on users' perceptions of deception. In *Proceedings of SPIE - Human Vision and Electronic Imaging XIX*, volume 9014, 2014.

[7] W. Li, Y. Yuan, and N. Yu. Passive Detection of Doctored JPEG Image via Block Artifact Grid Extraction. *Signal Process.*, 89(9):1821–1829, Sept. 2009.

[8] S. E. Middleton. Extracting Attributed Verification and Debunking Reports from Social Media : MediaEval-2015 Trust and Credibility Analysis of Image and Video. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[9] C. Pasquini, C. Brunetta, A. F. Vinci, V. Conotter, and G. Boato. Towards the verification of image integrity in online news. In *Proceedings of Multimedia Expo Workshops (ICMEW)*, pages 1–6, June 2015.

[10] Z. Zhao, P. Resnick, and Q. Mei. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405, 2015.