MCG-ICT at MediaEval 2016: Verifying Tweets From Both Text and Visual Content

Juan Cao[1], Zhiwei Jin[1,2], Yazi Zhang[1,2], Yongdong Zhang[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{jinzhiwei, caojuan, zhangyazi, zhyd }@ict.ac.cn

ABSTRACT

The Verifying Multimedia Use Task aims to automatically detect manipulated and fake web multimedia content. We have two important improvements this year: On the one hand, considering that the prediction based on a short tweet is unreliable, we propose a topic-level credibility prediction framework. This framework exploits the internal relations of tweets belonging to same topic. Besides, we enhance the prediction precision of the framework by sampling topics and exploring topic-level features. On the other hand, motivated by the idea that manually edited or low-quality videos tend to be fake, we reference the handbook[1] about detecting manual editions and build a decision tree on videos.

1. PROPOSED APPROACH

We treat the task as a binary classification problem: real or fake. Generally, a tweet contains two kinds of content: text content and visual content. So, we build two classification models respectively: for text content, the task pays more attention to small events than breaking news this year. More than 59% events contain less than 10 tweets and 95% are less than 50 tweets. Compared with last year's 42.5 tweets per event, the small event verification is more challenging. We propose a topic-level verification framework, and improve its performance by exploring topic-level features and sampling on topics. For visual content, we reference the handbook[1] about detecting manual editions and build a decision tree on videos.

However, the task focuses only on detecting fake tweets while we put efforts to identify on both categories. Finally, we propose a method performing relatively well on both real and fake tweets. More details about the task can be found in [2].

1.1 Text Analysis Approach

As illustrated in Figure 1, the framework of our text analysis approach consists of three parts: a message-level classifier, a topic-level classifier and a fusing part. Like many traditional text analysis methods, we firstly build a message-level classifier based on the given content features and user features. However, a tweet is very short (no more than 140 words) and its meaning is incomplete. The credibility prediction on message-level is unreliable.

We observe that each tweet contains videos/images in our data, and tweets containing the same videos/images are rather independent but have strong relations with each other. More specifically, they attend to have same credibilities. In order to exploit their inner-relations, we take the tweets of the same video/image as a topic, and build a topic-level classifier. Compared with an independent tweet, a topic can maintain principal information and also eliminate random noise. As the primary contribution in the text analysis, the topic-level improves the F1 value of 4% on fake tweets and more than 8% on real tweets. Two main innovations of the topic-level part are as follows:

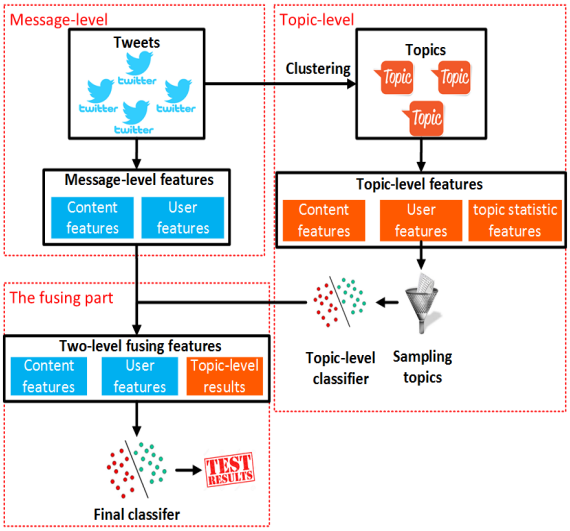

Figure 1: The framework of the text analysis approach

Topic-level Features Extracting: For each topic, we compute the average of its tweets' features as its features. Besides, we propose several statistic features which are listed in Table 1. Combining the two kinds of features above, we finally get the whole topic-level features. It turns out that these statistic features are quite effective for identifying fake tweets. They boost the topic-level classifier's F1 value on fake tweets by more than 14%.

Topics Sampling: In our dataset, More than 59% topics contain less than 10 tweets and 95% are less than 50 tweets, which means there are quite a few small topics. To remove the noise brought by these small topics, we sample topics with high confidence in a 10-fold cross validation process. The sampling keeps the balance between fake and real

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

Table 1: Topic Layer New Statistic Features

Feature	Explanation
num_tweets	the number of tweets
num_distinct_tweets/ hashtags	the number of distinct tweets /hashtags
distinct_tweets_index	the ratio of distinct tweets
contain_url/mention	the ratio of tweets containing urls/metions(namely, @)
contain_urls/mentions/ hashtags/questionmarks	the ratio of tweets containing multiple urls/mentions /hashtags/questionmarks

topics. By this technology, we largely improve the model's performance on real tweets.

The topic-level classifier classifies each topic and gives a corresponding probability value indicating how likely it is to be fake. In the fusing part, this probability, as the topic-level result, is added to its tweets' message-level features. The final classifier is built on the fused features above.

1.2 Visual Analysis Approach

In the testset, nearly half tweets contain videos while the other half contains images. Observing this, we build two visual classifiers respectively: for those tweets containing images, we propose the given 7 types of forensics features [3] [4] [5] [6] to build a image classifier model. For tweets containing videos, we build a decision tree which is the primary innovation of the visual analysis approach. Details about the tree are as follows:

The basic principle is that low-quality and manually edited videos(except the professional-edited news videos) are more likely to be fake. To detect manual editions, we reference the handbook[1] written by experienced journalists on recognizing manipulated videos, and summarize several features indicating whether the video is edited. The features include logos, video length, video size, shot number, resolution ratio, contrast ratio. Basing on these features, we build a decision tree which is illustrated in Figure 2.

The tree is intuitive. For a video containing logos, if it has a high quality, it's judged as professional-edited and the label is real; if it has a low quality, its label depends on the length: if it's long, it's more likely to be the kind of videos which is produced by original people and edited by professional journalists. So, the label is real, or otherwise it's fake. For a video lacking logos, if it has many shots which also suggests manual editions, it's fake. Otherwise it's real. More details about these features are as follows:

Logo Detecting: The basic idea to detect logos is that they are invariant compared with other parts of videos: we divide videos into frames, and detect color-fixed pixels in each frame. If a certain area's pixels keeps unchange for most frames, it is determined as a logo. To reduce random errors caused by steady dispersed pixels like short lines, we perform a median filter and we only keep logos that pass the filter.

Quality: We use the average value of video size, resolution ratio and contrast ratio to represent a video's quality.

Our video classification model reaches a F1 score of 0.702 on real tweets and 0.429 on fake tweets. However, there's a super video which contains 334 tweets while the other 25 videos contains only 777 in total. This super video brings

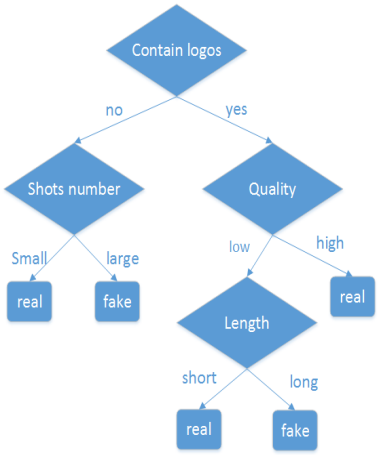

Figure 2: The video classification decision tree.

quite variations. Ignoring this video, our model reaches a F1 score of 0.918 on real tweets and 0.763 on fake tweets.

2. RESULTS AND DISCUSSION

We submitted 3 results which are listed in Table 2. Run 1 only uses text analysis approach while Run 2 only uses visual analysis approach. Run 3 is a hybrid of text and visual approach: if a sample of testing tweet contains videos/images, we use the visual model to classify it, otherwise we choose the text model.

Table 2: Topic Layer New Statistic Features

	Recall	Precision	F1-Score
Run 1	0.629	0.747	0.683
Run 2	0.514	0.698	0.592
Run 3	0.610	0.764	0.678

From the results we can observe that: (1)Both two models reaches very promising results. (2) The text model is more effective than the visual model. We assume it's probably because of lacking sufficient videos and our video model is under-fitting. But the idea to detect manual editions in videos inspires us to explore more videos to validate and improve our model in the future.

3. ACKNOWLEDGMENTS

This work was supported by National Nature Science Foundation of China (61571424, 61172153) and the National High Technology Research and Development Program of China (2014AA015202).

4. REFERENCES

[1] Craig Silverman. *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*. Number 121. The European Journalism Centre, 2014.

[2] Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, StuartE. Middleton, Andreas Petlund, and Yiannis Kompatsiaris. Verifying multimedia use at mediaeval

2016. In *Proceedings of the MediaEval 2016 Workshop*, Hilversum, Netherlands,Oct. 20-21, 2016.

[3] C.Pasquini, F.Perez-Gonzalez, and G. Boato. A benford-fourier jpeg compression detector. In *IEEE The International Conference on Image Processing, pp*, pages 5322–5326. IEEE, 2014.

[4] T.Bianchi and A.Piva. Image forgery localization via block-grained analysis of jpeg artifacts. In *IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, 2012.*, pages 1003–1017. IEEE, 2012.

[5] M.Goljan, J.Fridrich, and M.Chen. Defending against fingerprint-copy attack in sensor-based camera identification. In *IEEE Transactions on Information Security and Forensics, vol. 6, no. 1, 2010.*, pages 227–236. IEEE, 2010.

[6] W.Li, Y.Yuan, and N.Yu. Passive detection of doctored jpeg image via block artifact grid extraction. In *ACM Signal Processing, vol. 89, no. 9, 2009.*, pages 1821–1829. IEEE, 2009.