MCG-ICT at MediaEval 2015: Verifying Multimedia Use with a Two-Level Classification Model

Zhiwei Jin[1,2], Juan Cao[1], 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 detect misuse of online multimedia content and verify them as real or fake. This is a highly challenging problem because of strong variations among tweets from different events. Traditional approaches train the classifier at message level, which ignores inter-message relations. We propose a two-level classification model to exploit the information that tweets of a same topic are probably have same credibility values. In this model a topic level is introduced to eliminate message variations. Messages are aggregated into topics as a higher level representation. Pre-results gained from classification at the topic level are then fused with original message level features to train a better classifier. Results indicate that topic level is very helpful and our two-level approach offers significantly better results than a traditional one-level method. Our best result on this task achieves an F-score of 0.94 using features extracted only from tweet content.

1. PROPOSED APPROACH

The paper presents the approach developed by MCG-ICT for the MediaEval 2015 Verification Multimedia Use task. The task deals with the automatic detection of manipulation and misuse of Web multimedia content. Online content verification is a fairly new problem, participants are encouraged to propose effective features and methods. The goal of the task is to evaluate a set of tweets from several events and identify them as real or fake. More details about the task can be found in [1].

1.1 Two-Level Classification Model

Traditional approaches formulate the verification problem as a two-class classification task [2]. Features from tweet text contents and users are extracted to train a classifier at the message (tweet) level. One problem of this training strategy is that tweets are trained and tested individually. However, tweets in reality have strong relations among each other, especially tweets of a same topic would probably have the same verification result: real or fake.

Rather than classifying each tweet individually, some recent studies propose to verify tweets as a whole with inter-tweets information. Gupta et al. [3] propose a network which consists of tweets and users with similarity links among them. In our recent work [4, 5], we cluster tweets into sub-events and build links among tweets, sub-events and event. The three-layer network captures entities' relations from different scales and results in good verification performance.

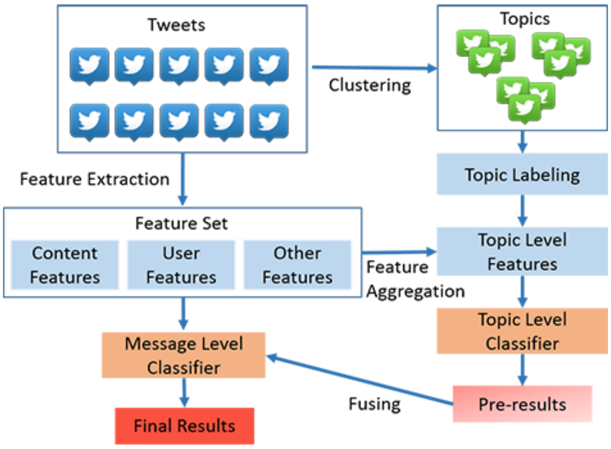

Figure 1: The framework of proposed two-level classification model. Topic level classification results are fused with message level to produce a final result.

Our network model is designed to evaluate the credibility of a specific event. However, in the presented data set of the target task [1], some events are actually a set of many related events (e.g. Hurricane Sandy) while some events only contain a few tweets (e.g. Pig Fish). Moreover, the task aims to give each tweet a verification label rather than an over all verification label for the event. These differences in the dataset and task definitions limit our model to directly work on it. But the idea of exploiting inter-tweet implications inspire us to propose a two level classification method. Figure 1 gives an overview of this method.

As illustrated in Figure 1, the proposed model has two levels of classifications: One is the message level which is just the same as previous message level methods. Features extracted from tweets text content, user information and other aspects are used for training; the other is the topic level which is the main contribution of this paper. By assuming tweets under a same topic probably have similar credibility values, we cluster tweets into different topics. A topic is a specific subject in an event, it consists of all tweets concerning the same subject. Compared with raw tweets, topics eliminate variations of tweets by taking the average of

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

them. Thus, it also reduces the impact of noisy data. Compared with event, topics maintain most of tweet details.

Topics Clustering: In [4], a clustering algorithm is used to cluster tweets into sub-events. But this algorithm performs poorly in forming topics in the task dataset as it is difficult to decide the optimum number of clusters. However, we observe that each tweet contains an image or video and each image or video can be contained in more than one tweets. This intrinsic one-to-many relations in the data set is a clue to form topics. To form topics, each image/video corresponds to a topic and tweets containing this image/video belong to this topic.

Topics Labeling: We label each topic as the average labels of its tweets: if more than a half of tweets in a topic is real then we label this topic as real. The labels are used for training the topic level classifier. (In fact, with the proposed topic formation, almost all tweets in a topic have the same label.)

Topic Level Feature Aggregation: We take the average of message level features of all tweets in a topic as the topic level feature. Some nominal features, such as ”contains question/exclaimation mark”, are also aggregated into corresponding numeric features.

Fusing Topic Level Result: After topic level classification, we can get a probability value for each topic on predicting how likely it is fake. Then for each tweet in the topic, we add this pre-result value as a feature to its original feature vector. Finally, we train a message level classifier with extended message features and give the final result.

1.2 Feature Set

In [2], 18 content features and 7 user features are extracted from the message level. We use these two kinds of features as base features. In addition, we also experiment on some new features: word term features and several image features.

We extract the commonly used term frequency (tf) features and tf-idf features to represent each tweet. With experiments on the training set (development set), this kind of feature was found to be very over-fitting. It reached very high performance on cross validation and very low performance on event-separation validation. Because few words co-occur in different events, we assume other pure term-based features (e.g. LDA features) would contribute little on this task.

Almost each tweet contains an image in the dataset, so we extract several features concerning images (e.g. image popularity, resolution). These image features can replace the topic level features to train classifier at topic level, because a topic is generated for each image as mentioned earlier. Experiments on the development set show that these image features result in slightly worse performance for the topic level classification than content features but much worse performance after fusing with message level features to generate the final result. Moreover, these image features cannot be applied directly on videos included in the test set. As they are not the main concern of this paper, we leave these features to future research.

2. RESULTS AND DISCUSSION

In the task requirements definition, runs 3-5 are experiments with external resources. As our approach focuses on the classification method rather than using external materials, we only submitted results for the first two runs (Table 1). Run 1 uses only content features while run 2 uses both content and user features. Both runs follow our two level classification model illustrated in Figure 1. We use J48 decision tree classifier for topic level classification and Random Forest classifier for message level classification. The topic level classification for training set is built by a 10-fold cross validation on it. The reported three evaluation measures in Table 1 are computed with respect to fake tweets.

Table 1: Verifying Multimedia Use Results

	Run 1	**Run 2**
Recall	0.9212	0.9220
Precision	0.9645	0.9374
F-Score	0.9423	0.9296

From the results we can observe that our two level classification method achieves very promising results on both runs. Specifically, it reaches a verification F-Score of 0.9423 for run 1 and a slightly worse result 0.9296 for run 2. Moreover, our method achieves high recall performance as well as high precision. This demonstrates the strong distinctive ability of our method for both fake and real tweets. We also notice that the result of run 2 is slightly worse than that of run 1, which indicates the user features may be redundant. In fact, we get a similar result in our experiments on the development set.

In the future, we want to explore other features, such as image forensics features, with our model. This model also need to be tested on a much larger data set or in real-time situations to validate its effectiveness.

3. ACKNOWLEDGMENTS

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

4. 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 *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, 2015.

[2] C. Boididou, S. Papadopoulos, and Y. Kompatsiaris. 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.

[3] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the SIAM International Conference on Data Mining*, page 153. Society for Industrial and Applied Mathematics, 2012.

[4] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 230–239. IEEE, 2014.

[5] X. Zhou, J. Cao, Z. Jin, X. Fei, Y. Su, J. Zhang, D. Chu, and X. Cao. Real-time news certification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 983–988, 2015.