The CERTH-UNITN Participation @ Verifying Multimedia Use 2015

Christina Boididou[1], Symeon Papadopoulos[1], Duc-Tien Dang-Nguyen[2], Giulia Boato[2], 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

ABSTRACT

We propose an approach that predicts whether a tweet, which is accompanied by multimedia content (image/video), is trustworthy or deceptive. We test different combinations of quality and trust-oriented features (tweet-based, user-based and forensics) in tandem with a standard classification and an agreement-retraining technique, with the goal of predicting the most likely label (`fake` or `real`) for each tweet. The experiments carried out on the Verifying Multimedia Use dataset show that the best performance is achieved when using all available features in combination with the agreement-retraining method.

1. INTRODUCTION

Since social media have gained momentum over the years as a fast and real-time means of sharing news, a huge amount of information is constantly flowing through it, quickly reaching massive numbers of readers. Thus, it can easily become viral and affect public opinion and sentiment. This has motivated a number of malicious efforts to spread misleading content, highlighting the need for fast verification. In this setting, the goal of Verifying Multimedia Use task is to automatically predict whether a tweet that shares multimedia content is misleading (referred to as `fake`) or trustworthy (`real`) [1]. To this end, we make use of the tweet text content, a set of tweet- and user-based features and multimedia forensic features for the images embedded in the tweet.

In our work, we present an extension of our original approach [2], combining different sets of the aforementioned features. The conducted experiments include plain classification models and an agreement-retraining method that uses part of its own predictions as new training samples with the goal of adapting to the new event. In the next sections, we present in detail the adopted methodology.

2. SYSTEM OVERVIEW

2.1 Features

The approach uses three types of features: a) tweet-based (`TB`), which make use of information coming from the tweet and its metadata, b) user-based (`UB`), which are computed using information and metadata about the user posting (or retweeting) the tweet, c) multimedia forensics features, which are computed based on the image that accompanies the tweet. We test two variants of the first two sets of features: i) baseline (`base`), which correspond to the features shared by the organisers, and ii) extended (`ext`), which include a few new features. The forensics features include both the ones distributed by the organisers and some additional ones.

Table 1: List of features used in the experiments.

Feature set	Description
`TB-base`	Baseline tweet-based
`TB-ext`	Extended tweet-based
`UB-base`	Baseline user-based
`UB-ext`	Extended user-based
`FOR`	Forensic features

`TB-ext`: We extract additional features based on the tweet text, such as the presence of a word, symbol or external link. We also use language-specific binary features that correspond to the presence of specific terms; for languages, in which we cannot manage to define such terms, we consider the values of these features missing. We perform language detection with a publicly available library[1]. We add a feature for the `number of slang words` in a text, using slang lists in English[2] and Spanish[3]. For the `number of nouns`, we use the Stanford parser[4] to assign parts of speech to each word (supported only in English). For the readability of the text, we use the Flesch Reading Ease method[5], which computes the complexity of a piece of text as a score in the interval $[0, 100]$ (0: hard-to-read, 100: easy-to-read).

`UB-ext`: We extract user-specific features such as the `number of media content`, the `account age` and others that refer to the information that the profile shares. For example, we check whether the user declares his/her geographic location and whether the location can be matched to a city name from the Geonames dataset[6].

Next, for both `TB` and `UB` features, we adopt trust-oriented features for the links shared, through the tweet itself (`TB`) or

[1]`https://code.google.com/p/language-detection/`

[2]`http://onlineslangdictionary.com/word-list/0-a/`

[3]`http://www.languagerealm.com/spanish/spanishslang.php`

[4]`http://nlp.stanford.edu/software/lex-parser.shtml`

[5]`http://simple.wikipedia.org/wiki/Flesch_Reading_Ease`

[6]`http://download.geonames.org/export/dump/cities1000.zip`

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

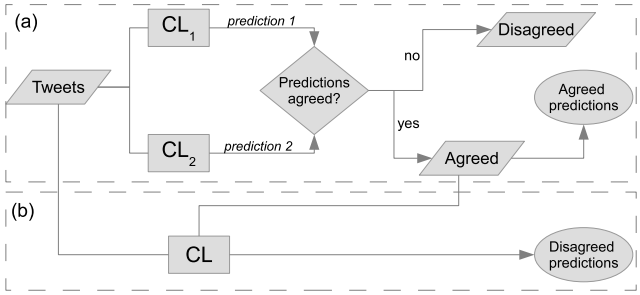

Figure 1: Overview of agreement-based retraining.

Table 2: Description of runs. For SSL-AR case, the two sets of features used for building the two classifiers are mentioned. For example, in RUN-3 the TB-base + FOR are used for CL_1 and the UB-base for CL_2.

Run	Features	Learning
RUN-1	SL	TB-base
RUN-2	SL	TB-base+FOR
RUN-3	SSL-AR	(TB-base+FOR) + UB-base
RUN-4	SL	TB-ext+UB-ext+FOR
RUN-5	SSL-AR	(TB-ext+FOR) + UB-ext

the user profile (UB). The WOT metric[7] is a score indicating how trustworthy a website is, using reputation ratings by Web users. We also include the in-degree and harmonic centralities, rankings computed based on the links of the web forming a graph[8]. Trust analysis of the links is also done using four Web metrics provided by the Alexa API[9].
FOR: For each image, the additional forensics features are extracted from the provided BAG feature based on the maps obtained from AJPG and NAJPG. First, a binary map is created by thresholding the AJPG map (we use 0.6 as the threshold), then the largest region is selected as *object* and the rest of the map is considered as the *background*. For both regions, seven descriptive statistics (maximum, minimum, mean, median, most frequent value, standard deviation, and variance) are computed from the BAG values and concatenated to a 14-dimensional vector. We apply the same process on the NAJPG map to obtain a second feature vector.

2.2 Agreement-based retraining method

The main extension of this system compared to [2] includes an *agreement-based retraining* step in order to improve the prediction accuracy for unseen events. This is motivated by a similar approach implemented in [3] (for the problem of polarity classification). Figure 1 illustrates the adopted process. In step (a), we build two classifiers CL_1, CL_2 based on the training set, each classifier built on different types of features, and we combine their outputs in a Semi-Supervised Learning (SSL) fashion. We compare the two predictions for each sample of the test set, and depending on their agreement, we divide the test set in two subsets, the *agreed* and *disagreed* samples. These two subsets are treated differently by the classification framework.

Assuming that the agreed predictions are correct with high likelihood, we use them as training samples to build a new classifier for classifying the disagreed samples. To this end, in step (b), we add the agreed samples to the best performing of the two initial models, CL_1, CL_2 (comparing them on the basis of their performance when doing cross-validation on the training set). The goal of this method is to retrain the initial model and make it adaptable to any specific characteristics of the new event. In that way, the model can predict more accurately the values of the samples for which CL_1, CL_2 did not agree in the first step.

[7] https://www.mywot.com/
[8] http://wwwranking.webdatacommons.org/more.html
[9] http://data.alexa.com/data?cli=10&dat=snbamz&url=google.gr

Table 3: Results.

	Recall	Precision	F-score
RUN-1	0.794	0.733	0.762
RUN-2	0.749	**0.994**	0.854
RUN-3	0.922	0.736	0.819
RUN-4	0.798	0.860	0.828
RUN-5	**0.969**	0.861	**0.911**

2.3 Bagging

Due to the unequal number of fake and real tweets, we exploit only a part of the data while building a model. In order to take advantage of the whole training dataset, we use bagging that tends to improve the accuracy of the method, as it produces predictions using the average result of numerous predictors. Bagging creates m different subsets of the training set, including equal number of samples for each class (some samples may appear in multiple subsets), leading to the creation of m instances of CL_1 and CL_2 classifiers ($m = 9$). The final prediction for each of the testing samples is calculated using the majority vote of the m predictions.

3. SUBMITTED RUNS AND RESULTS

The five runs submitted explore different combinations of features and the use of a standard supervised learning scheme (SL) versus the newly proposed agreement-based retraining (SSL-AR). The specific run configurations are specified in Table 2.

RUN-1, RUN-2 and RUN-4 are built using a plain classification model. RUN-3 and RUN-5 are built with the agreement-based retraining technique, in which we build CL_1 and CL_2 (Figure 1) by using the sets of features specified in Table 2. All models use a *Random Forest* classifier from the Weka implementation.

Table 3 presents the performance of each run. In terms of *F-score*, which is the primary evaluation metric of the task, RUN-5 achieved the best score when using the ext and the FOR features with the SSL-AR technique. As we observe, RUN-2 in which the FOR features are added, performed quite better than RUN-1, which uses just the TB-base features. Comparing RUN-4 and RUN-5, one may observe the considerable performance benefit stemming from the use of the SSL-AR approach, as it is the only difference between the two runs (the same sets of features are used). Additionally, it is important to note the contribution of the ext features, as RUN-5 (ext) performs better than RUN-3 (base).

4. ACKNOWLEDGEMENTS

This work is supported by the REVEAL project, partially funded by the European Commission (FP7-610928).

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

[2] 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.

[3] A. Tsakalidis, S. Papadopoulos, and I. Kompatsiaris. An ensemble model for cross-domain polarity classification on twitter. In *Web Information Systems Engineering–WISE 2014*, pages 168–177. Springer, 2014.