The VMU Participation @ Verifying Multimedia Use 2016

Christina Boididou[1], Stuart E. Middleton[5], Symeon Papadopoulos[1], Duc-Tien Dang-Nguyen[2,3], Michael Riegler[4], Giulia Boato[2], Andreas Petlund[4], 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

[3]Insight Centre for Data Analytics at Dublin City University, Ireland. duc-tien.dang-nguyen@dcu.ie

[4]Simula Research Laboratory, Norway. michael@simula.no,apetlund@ifi.uio.no

[5]University of Southampton IT Innovation Centre, UK. sem@it-innovation.soton.ac.uk

ABSTRACT

The participating approach predicts whether a tweet, which is accompanied by multimedia content (image/video), is trustworthy (`real`) or deceptive (`fake`). We combine two different methods a) one using a semi-supervised learning scheme that leverages the decisions of two independent classifiers to produce a decision and b) one using textual patterns to extract claims about whether a post is fake or real and attribution statements about the content source. The experiments, carried out on the Verifying Multimedia Use dataset, used different combinations of content quality and trust-oriented features, namely tweet-based, user-based and forensics.

1. INTRODUCTION

After high-impact events, large amounts of unverified information usually start spreading in social media. Often, misleading information is getting viral affecting public opinion and sentiment [4]. Based on this problem, the Verifying Multimedia Use task highlights the need for verification and addresses the challenging problem of establishing automated approaches to classify social media posts as containing misleading (`fake`) or trustworthy (`real`) content [2].

To tackle this challenge, we present a method combining two approaches. The first approach is an extension of [3] which introduces an agreement-retraining method that uses part of its own predictions as new training samples with the goal of adapting to posts from a new event (method `ARM`). The second approach uses textual patterns to extract claims about whether a post is fake or real and attribution statements about the source of the content [7] (method `ATT`). The conducted experiments use various sets of features.

2. SYSTEM DESCRIPTION

2.1 ARM: Agreement-based Retraining

Being an extension of [3], the proposed method uses an *agreement-based retraining* step with the aim to adapt to posts from new events and improve the prediction accuracy on them. This is motivated by a similar approach implemented in [8] (for the problem of polarity classification). Figure 1 illustrates the adopted process. In step (a), using the training set, we build two independent classifiers CL_1, CL_2 and we combine their predictions for the test set. We compare the two predictions, and depending on their agreement, we divide the test set into *agreed* and *disagreed* subsets, which are treated differently by the classification framework. Assuming that the agreed predictions are correct with high likelihood, we use them as training samples along with our initial training set to build a new model 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 make the model adaptable to specific characteristics of a new event.

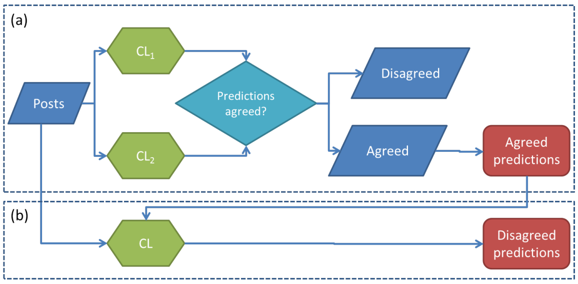

Figure 1: Overview of agreement-based retraining.

The classifiers are built using three types of features: a) tweet-based (`TB`), which use the post's metadata, b) user-based (`UB`), which use the user's metadata, c) multimedia forensics features (`FOR`), which are computed for the image that accompanies the post. Except for the ones shared by the task, we extract and use additional ones on each set.

`TB`: Binary features such as the presence of a word, symbol or external link are added to the list. 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 these values as missing. We perform language detection with a publicly available li-

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

brary[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) and for text readability, the Flesch Reading Ease method[5], which computes the complexity of a piece of text as a score in the interval (0: hard-to-read, 100: easy-to-read).

UB: We extract user-specific features such as the number of media items, the account age and others that summarize information shared by the user. For example, we check whether the user shares a location and whether this can be matched to a city name from the Geonames dataset[6].

For both TB and UB features, we adopt trust-oriented features for the links shared, through the post itself (TB) or 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 performed using Web metrics provided by the Alexa API.

FOR: Following the method in [1], forensics features are extracted as descriptive statistics (maximum, minimum, mean, median, most frequent value, standard deviation, and variance) computed from the BAG values. In this work, we also extracted an additional feature that can measure the image quality as a single score (from 0 to 100) by exploiting the method in [6]. The forensics features extraction step is performed as follows: for each image, a binary map is created by thresholding the AJPG map, then the largest region is considered as *object* and the rest as the *background*. For both regions, seven descriptive statistics are computed from the BAG values and concatenated to have a 14-dimensional vector. The same process is applied on the NAJPG map. In order to measure the image quality, discrete cosine transformation (DCT) is applied on the whole image, then a support vector machine is applied to predict the quality based on the values of the spectral and spatial entropies (computed from the block DCT coefficients). In the end, all the forensics features are concatenated as a 29-dimensional vector (14 from AJPG, 14 from NAJPG, and 1 from image quality).

2.2 ATT: Attribution based claim extraction

This approach is motivated by the human verification process employed by journalists, where attributed sources are key to trustworthiness of claims. A classic natural language processing pipeline is employed, involving text tokenization, Parts of Speech (POS) tagging[9] and a permissive named entity recognition pattern focussing on noun phrases. A number of regex patterns were created to extract typical linguistic constructs around image and video content, such as debunking reports, claims of being real or attribution to a third party source such as a news provider.

Our approach is semi-automated, using a list of a priori known trusted and untrusted sources. We can either learn an entity list automatically using information theoretic weightings (i.e., TF-IDF) or create a list manually (i.e., using a journalist's trusted source list). All news providers have long lists of trusted sources for different regions around the world so this information is readily available. For this task we created a list of candidate named entities by first running the regex patterns on the dataset. We then manually checked each entity via Google search (e.g., looking at Twitter profile pages) to determine if they were obvious news providers or journalists.

Table 1: Run description and results.

	Learning	Precision	Recall	F-score
RUN-1	ARM (TB,UB)	0.981	0.851	0.912
RUN-2	ARM (TB+FOR,UB)	0.771	0.906	0.833
RUN-3	ARM (TB,UB) and ATT (pr.)	**0.988**	0.887	**0.935**
RUN-4	ARM (TB,UB) and ATT (ret.)	0.980	0.874	0.924
RUN-5	TB,UB,FOR	0.587	**0.995**	0.739

We assign a confidence value to each matched pattern based on its source trustworthiness level. Evidence from trusted authors is more trusted than evidence attributed to other authors, which is more trusted than unattributed evidence. In a cross-check step we choose the most trustworthy claims to use for each image URI. If there is evidence for both a fake and genuine claim with an equal confidence we assume it is fake (i.e., any doubt = fake). Our approach provides a very high precision, low recall output.

3. SUBMITTED RUNS AND RESULTS

We submitted five runs that explore different combinations of features (TB, UB, FOR) and methods (ARM, ATT). Table 1 shows the specific run configurations and performance.

In RUN-1 and RUN-2, we apply the ARM in which we build CL_1 and CL_2 (Figure 1) by using the sets of features specified in Table 1. For example, in RUN-2, we use the concatenation of TB + FOR for CL_1 and UB for CL_2. RUN-3 is a combination of ARM and ATT methods, in which we consider for each post the result of ATT as correct if available, otherwise we use the output of ARM. Similarly, in RUN-4, we consider the results of ATT as samples for retraining (step (b) in Figure 1) along with the agreed ones of ARM. All models built in ARM use a *Random Forest* WEKA implementation [5]. Finally, RUN-5 is a plain classification method that is built with the whole amount of available features. In terms of performance (*F-score* is the evaluation metric of the task), RUN-3 achieved the best score when using the combination of the two methods. Apparently, as shown from the RUN-1, RUN-2, the presence of FOR features reduced the system's performance. By observing the RUN-3 and RUN-4, one may notice the considerable performance benefit is derived from the combined use of ARM and ATT.

4. ACKNOWLEDGMENTS

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

5. REFERENCES

[1] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, and Y. Kompatsiaris. The certh-unitn participation@ verifying multimedia use 2015. *MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval'15)*, 2015.

[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

[7] https://www.mywot.com/

[8] http://wwwranking.webdatacommons.org/more.html

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[2] 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 *MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands*, 2016.

[3] 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.

[4] V. Conotter, D.-T. Dang-Nguyen, G. Boato, M. Menéndez, and M. Larson. Assessing the impact of image manipulation on users' perceptions of deception. *Proceedings of SPIE - The International Society for Optical Engineering*, 2014.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations*, 11(1):10–18, 2009.

[6] L. Liu, B. Liu, H. Huang, and A. Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014.

[7] S. E. Middleton. Extracting attributed verification and debunking reports from social media: Mediaeval-2015 trust and credibility analysis of image and video. 2015.

[8] 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.