Fusion-based multimodal detection of hoaxes in social networks

Cédric Maigrot, Vincent Claveau, Ewa Kijak

Fusion-based multimodal detection of hoaxes in social networks

Cédric Maigrot
Univ. Rennes, IRISA
Rennes, France
cedric.maigrot@irisa.fr

Vincent Claveau
CNRS, Univ. Rennes, IRISA
Rennes, France
vincent.claveau@irisa.fr

Ewa Kijak
Univ. Rennes, IRISA
Rennes, France
ewa.kijak@irisa.fr

***Abstract*—Social networks make it possible to share information rapidly and massively. Yet, one of their major drawback comes from the absence of verification of the piece of information, especially with viral messages. This is the issue addressed by the participants to the *Verification Multimedia Use* task of *Mediaeval 2016*. They used several approaches and clues from different modalities (text, image, social information). In this paper, we explore the interest of combining and merging these approaches in order to evaluate the predictive power of each modality and to make the most of their potential complementarity.**

***Index Terms*—Hoax detection, Knowledge fusion, Text analysis, Image analysis, Source credibility.**

I. Introduction

Social networks are becoming increasingly important in our professional and personal lives, thanks to their ability to keep us informed through the news shared by our contacts. It has become common that important news is first broadcast on social networks before being processed by traditional media. This speed of the propagation of information, combined with the number of people receiving it, defines the virality of information. But this virality has a drawback: users rarely check the veracity of the information they share. It is therefore common to see false and/or manipulated information circulating (including hoaxes, rumors, urban legends, or fake news). In addition, even information identified as a hoax can be difficult to stop when it is already shared a large number of times.

The purpose of this project is to automatically detect the veracity of viral information. The ultimate goal is to create for example a system that will warn the user before he shares false information. Starting from the observation that this viral information is often composed of multimedia elements (text accompanied by images or videos), we propose to develop systems exploiting different modalities. We present first approaches exploiting only the textual content, the contents of the images or the sources cited in the messages, and then different strategies of combination of these mono-modal approaches. These different approaches are evaluated and discussed on the shared data of the *Verifying Multimedia Use* (VMU) task from the challenge *MediaEval2016* that was specifically on this issue. From the methods of all the teams involved in this task, we explore different fusion strategies to analyze the contribution of different approaches and the predictive ability of a collaborative system.

After a review of the state of the art in the next section, we present in Section III the task *VMU* and its datasets. We then present in Section IV the approaches we have implemented, as well as the systems proposed by the other teams participating in the *VMU* task. Sec. V presents the experimental protocol and the results obtained by the different approaches. Different fusion strategies are tested and discussed in Sec. VI. Finally, Sec. VII summarizes the main observations and discusses possible avenues for the future.

II. State of the art

The analysis of the information veracity is a research challenge that is studied from several angles. Here, we are only interested in viral information circulating in social networks (SN). It should be noted that research about *fact checking* is not discussed here; even if it shares a common goal of verification, the differences in the nature of information (source, mode of dissemination) and purpose (help for journalism) imply different methods from those used for hoax.

Several families of features have been exploited to detect *hoax* in social networks. We can notably cite:

- the textual indices; the message itself obviously brings potentially exploitable information.
- multimedia cues, in the case of messages containing images or videos; these media contents can sometimes be analysed to detect some deliberate modifications (*tampering*).
- the diffusion in the SN: what is the source of the information, what is the path of the message.

Analysis of the sources of messages and relations between members has been the subject of several studies. [1] offer a measure of trust between users of SN, which characterizes the trust of a relationship between two users. This measure of trust can thus serve as an index to judge the reliability of the information transmitted. Several approaches have also been proposed to determine the credibility of a source. [2] adapt the *PageRank* algorithm [3] on a graph representing the relationships between tweets, the authors of these tweets and the events associated with these tweets. These approaches, however, require extensive knowledge of the network that makes them difficult to apply in practice for large-scale commercial SN.

The analysis of the mode of dissemination of the messages as well has been the subject of several works, the purpose of

which is to distinguish rumors from conventional messages by observing how they propagate in the SN. These analyzes are based on diffusion models but require access to a large part of the network to follow these messages, which is rarely possible with the mainstream SN we are aiming for.

Analysis of multimedia clues is at the heart of the European project *InVid*[1] which focuses on the automatic detection of fake videos in SN. For this purpose, [4] work on images from videos and analyze them with *forensics* techniques to detect modifications in the images. The problem of image authenticity has been addressed in the *Image forensics* community in several ways [5], [6]. However most of the approaches poorly perform on images circulating on SN [7]. In this context, other approaches exploits external information to determine the integrity of an image. In this case, similar or identical images are searched in database (or Web) in order to determine if the image has been modified or diverted. The problem of finding similar images is an active domain whose latest work is based on deep convolutional neural networks to describe and compare images [8].

Other works exploit several of these indices together. This is the case of [9], which, within the framework of the European *Reveal Project* [2], aims to develop tools and services for information verification in social networks, from a journalistic and professional perspective. Different media such as image [10], video [11], and text are analyzed. However, their purpose is not to develop automatic tools but to help journalists. These works also extensively use external resources [12].

The European project *PHEME*[3] [13] is interested in detecting rumors in SN or online media. In particular, they study the responses and reactions to tweets to decide of their truth. This project is not intended, as we do in this article, to classify the tweet based on its unique content, but instead, aims to *crowdsourced verification*, that is to say relies on (human) analyzes produced by the users of the SN.

III. *Verifying Multimedia Use* TASK FROM *MediaEval2016*

The *Verifying Multimedia Use* (VMU) task of the 2016 MediaEval evaluation campaign proposed to classify messages from *Twitter* according to their veracity (classes *true* or *false*, with the possibility of using an *unknown* class if the system is not able to take a decision). Allowing the system not to decide can result in high precision for the *true* and *false* classes [14].

In the provided evaluation data, all messages are labeled either *true* or *false*, and all are accompanied by one or more images, or a video (see Fig. 1). However, several messages may share the same image. It is important to note all messages sharing the same video or a image will always have the same class (creating some bias resulting from the dataset building method). While some images are only used by a single message, others are shared by more than 200 messages. In addition, messages are grouped by event. The sizes of the events are not balanced. For instance, the biggest event in this collection is *Paris Attack* with 580 messages sharing 25 different images or videos, while the smaller ones are the events *Soldier Stealing* and *Ukrainian Nazi* with one single message and one single image. Tab. I shows the distribution of data between learning and test sets, as well as the number of images and videos per set. It should be noted that in Sec. V, the results presented are those obtained on the test set, and that the fusion techniques presented in Sec. VI are used on the predictions of participants' systems on this test set (*i.e.* the submitted runs).

[1]See http://www.invid-project.eu/

[2]See https://revealproject.eu

[3]See https://www.pheme.eu/

Fig. 1. Examples of two tweets from the MediaEval VMU task, on the same event (*Hurricane Sandy*)

Training set		Test set	
17 events	15,821 messages	35 events	2,228 messages
True	False	True	False
6,225 mes.	9,596 mes.	998 mes.	1,230 mes.
Images 193	Images 118	Images 54	Images 50
Videos 0	Videos 2	Videos 10	Videos 16

TABLE I
VMU TRAINING AND TEST SET DESCRIPTION

Several features have been proposed by the organizers, falling into three categories: textual, user or image. The proposed textual features, noted $\mathcal{T}$, are shallow ones: number of words, length of the text, occurrence of the symbols *?* and *!*, presence of the symbols *?* and *!* as well as happy or unhappy emoticons, pronouns at the 1st, 2nd or 3rd person, number of capital letters, number of positive and negative words, number of mentions *Twitter*, hashtags, urls and retweets.

The set of feature associated with the user, noted $\mathcal{U}$, consists of the following information: number of friends, number of subscribers (*followers*), ratio of the number of friends on the number of subscribers, if the account contains an url, if the account is verified and the number of messages posted.

The set of features associated with the images, noted $\mathcal{FOR}$, comes from the domain of *forensics*: indexes of double JPEG compression [15], Block Artifact Grid [16], Photo Response Non-Uniformity [17] and Benford-Fourier coefficients [18].

IV. SYSTEMS PARTICIPATING TO MEDIAEVAL VMU TASK

Four teams participated in the task for a total of 14 bids. Teams are subsequently denoted by *LK* (our team), *MMLAB*,

MCG-ICT and *VMU* (task organizers). In this section, we present the approaches we have developed, then briefly the approaches proposed by the other participating teams.

In our approaches, all messages sharing the same image are associated with the same *true*, *false*, or *unknown* class. It is thus sufficient to determine the class of each image and assign the predicted class to all messages associated with this image, according to the following rule: a message is predicted as *true* if all the associated images are classified *true*; it is predicted *false* otherwise. We propose three approaches: the first is based on the textual content of the message; the second on sources; the third on the images. None of our approaches uses the descriptors $\mathcal{T}$,$\mathcal{U}$ or $\mathcal{FOR}$ presented in Sec. III, but other participants do. We present their approaches at the end of this section. A comparative study between all approaches will be proposed later in the article.

A. Text-based run (LK-T)

This approach exploits the textual content of messages and does not rely on additional external knowledge. As previously explained, a tweet is classified from the associated image. An image is itself described by the union of the textual contents of the messages that use this image. The idea behind this approach is to capture similar comments between a test set message and those of the training set (*e.g.* "it's photoshopped") or more stylistic aspects (*e.g.* presence of emoticons, popular expressions...).

Let I_q be the description of an unknown image (*i.e.* the union of the textual contents of all the messages that use this image), and $\{I_{d_i}\}$ be the set of descriptions of the images in the training set. The class of I_q is determined by a k-nearest neighbors classification (majority vote). The calculation of similarity between textual descriptions is therefore at the heart of this approach, as the k images in $\{I_{d_i}\}$ whose descriptions are the most similar to I_q should be identified. The similarity used is Okapi-BM25 [19].

B. Source-based run (LK-S)

This approach, similar to [11], is based on external knowledge (static). As for the previous approach, the prediction is done at the image level, and the image is represented by the union of the textual contents (translated in English if necessary) of the messages in which it appears. The prediction is made by detecting a source of confidence in the description of the image. Two types of sources are sought: 1) a known information organization; 2) an explicit quote of the source of the image. For the first type of source, we determine a list of news agencies and newspapers in the world (mainly French and English) based on established lists[4], and information television networks (French and English)[5]. For the second type, we manually define several extraction patterns, like `photographed by + Name`, `captured by + Name`, ...Finally, an image is classified as *unknown* by default, except if a source of confidence is found in his description.

[4]https://en.wikipedia.org/wiki/List_of_news_agencies

[5]https://en.wikipedia.org/wiki/Lists_of_television_channels

C. Image-based runs (LK-I and LK-I2)

In this approach, only the content of the images is used to make a prediction. Tweets containing videos are not handled by this approach and get the *unknown* class. It relies on a similar-image search in a database of reference images, listed as *false* or *true*. A given query image (which we are looking for the class) receives the class of the most similar image of the database (if it exists), otherwise query image and associated messages are labeled *unknown*.

The image database was built by collecting images from five websites specialized in debunking false information: *www.hoaxbuster.com*, *hoax-busters.org*, *urbanlegends.about.com*, *snopes.com* and *www.hoax-slayer.com*. The database contains about 500 original images (*i.e. true*) and 7,500 tampered images (*i.e. false*).

The descriptors we generate from images are calculated using a deep convolutional neural network (CNN) [20]. The images are first resized to the standard size of 544×544 and passed in the CNN layers [21] to get a $4,096$ dimension description vector. Once the image descriptors are obtained, a cosine similarity is calculated between the query images and the images of the database. The search system therefore returns a list of images ordered by similarity. To consider that two images are sufficiently similar, their cosine must exceed 0.9 (determined empirically on the training set).

In the `LK-I` approach, if no image of the database is found to be similar, the query image receives the *unknown* class. Due to the small size of the reference base, this case is common. An alternative version of this approach, noted `LK-I2`, assigns to these uncertain images the maximum prior probability class that is *false*.

D. Presentation of other teams' runs

For each of the other participating teams, we describe below the data and method used to predict the class of messages.

1) VMU *runs:* Five methods were tested by the task organizers. These methods are based on two systems, of which they are variants [22].

`VMU-F1` and `VMU-F2` rely on a first system which is a meta-classifier in which two sets of descriptors are used separately by two classifiers, trained on the training set. Each classifier then predicts *true* or *false* for each message, which makes it possible to obtain two predictions per message. Messages from the test set that have received different predictions are further processed by a third classifier (Random Forest) trained on the union of the training set and messages of the test set having received predictions in agreement on both first classifiers. `VMU-F1` uses the $\mathcal{T}$ and $\mathcal{U}$ descriptors for the first two classifiers, while `VMU-F2` uses the union of $\mathcal{T}$ and $\mathcal{FOR}$ for one of the classifiers, and $\mathcal{U}$ for the other.

`VMU-S1` and `VMU-S2` are based on `VMU-F1`, to which is added a second system which exploits two lists of known sources: the first is a list of sources of trust while the

second groups sources of non-confidence. When a source-based prediction is not possible, the first system is used to provide a prediction. Finally, VMU−B is a reference obtained by the application of a classifier on the concatenation of $\mathcal{T}$, $\mathcal{U}$ and $\mathcal{FOR}$ descriptors.

2) MMLAB *runs:* The proposed approach is based on two random forest classifiers [23]. The first classifier, called MML−T, takes as input the concatenation of the $\mathcal{T}$ and $\mathcal{U}$ descriptors proposed by the task organizers.

The second classifier, denoted MML−I, uses the multimedia contents (images and videos) associated with the messages. It takes as input the concatenation of forensics descriptors (the set $\mathcal{FOR}$) and textual descriptors obtained using an external knowledge base. For each event, using the *TF-IDF* metric on the texts, a list of the most relevant terms related to the event is established from the most relevant web sites returned by an online text search engine. For each image, an inverted search engine (*Google image search*) is then used and frequency measurements of (i) previously identified relevant terms, (ii) positive and negative polarity terms (derived from the lexicon used in sentiment analysis) are applied to the texts of the most relevant sites found. In the case of a video coming from *Youtube*, these frequency measurements are applied to the comments of the video. The other videos are not analyzed.

Finally MML−F is the fusion (linear combination) of the scores of each class provided by MML−T and MML−I with respective coefficients of 0.2 and 0.8 in order to favor the second module while ensuring a prediction in the case of prediction failure of the second module (*e.g.* video not coming from *Youtube*).

3) MCG-ICT *runs:* The first approach proposed by [24] is based on the text content of messages. The $\mathcal{T}$ and $\mathcal{U}$ descriptors are used, and a new descriptor is added to this set. The computation of this new descriptor is based on the separation of an event into themes, a theme being defined as the set of messages sharing the same image or video. Each theme is described by the average of the $\mathcal{T}$ and $\mathcal{U}$ descriptors of its messages, and by statistics such as the number of messages in the theme, the number of distinct messages (*i.e. hashtags*, to discriminate retweets), the ratios of distinct messages, of messages containing a URL or mention, and of messages containing multiple URLs, mentions, hashtags or question marks. From these characteristics, a theme-level classifier is constructed, and indicates the probability of a message to be *true* or *false*. This probability is the new descriptor added to each message. The classifier at the message level, built on the enriched textual descriptors, is called MCG−T.

A second module evaluates the credibility of the visual content. For images, authors use the $\mathcal{FOR}$ descriptors (without specifying the classifier used). Videos are treated differently. Referring to [25], the authors define four characteristics to describe videos: a measure of the sharpness of the image, the contrast ratio, defined as the ratio of the size of a video over its duration, the duration of the video and the presence of logos. These four characteristics are combined by a binary decision tree. The predictions corresponding to this approach are MCG−I.

Finally, MCG−F combines these two previous predictions.

	F-score	Accuracy
LK−T	71.7 (36.9)	69.5 (36.9)
LK−I	47.5 (45.6)	45.8 (45.6)
LK−I2	80.7 (33.5)	78.8 (35.0)
LK−S	**81.9 (33.8)**	**84.3 (30.6)**
VMU−F1	28.9 (39.7)	40.8 (39.0)
VMU−F2	71.1 (40.4)	**74.4 (36.4)**
VMU−S1	40.0 (43.8)	50.9 (41.1)
VMU−S2	33.5 (41.2)	43.6 (40.3)
VMU−B	**77.2 (33.7)**	74.1 (35.2)
MML−T	9.25 (23.3)	13.8 (23.9)
MML−I	70.4 (36.4)	67.3 (36.4)
MML−F	**71.3 (36.7)**	**71.5 (34.3)**
MCG−T	**66.4 (42.9)**	**67.5 (41.3)**
MCG−I	55.9 (42.9)	59.9 (40.5)
MCG−F	62.6 (43.4)	66.6 (41.0)

TABLE II
PERFORMANCE OF RUNS SUBMITTED TO THE VMU TASK IN TERMS OF ACCURACY (%) AND MICRO-F-SCORE (%) (STD-DEV IN PARENTHESIS)

V. RESULTS OF THE SUBMITTED RUNS

A. Experimental setup

The data used to evaluate these systems are those from the test set of the *VMU* task presented in Sec. III. The evaluation measure used in the task was the *F-score* on the *false* class. However, this measure is not discriminating between the predictions *true* and *unknown*, and is based on the majority class *false*, which represents a bias: a system predicting *false* for every message would get 71.14% as F-score. We thus use instead the micro-F-score and the accuracy which are global measurements on all the classes to be predicted.

In addition, as an image that can be used by several messages, the evaluation is done by cross validation on the events, in order to guarantee that all the messages using the same image are in the same fold so that the assessment is not biased. To implement this cross-validation, all events are randomly subdivided into n packets. The evaluation therefore accounts for the performance that can be expected when processing a new event generating a set of messages that can be true or false. The results of the methods described in Sec. IV, re-evaluated according to the protocol described above (scores and evaluation by sets of messages sharing the same image), are presented in Tab. II.

Between the two evaluation modes (per message, or group of messages sharing the same multimedia content), there are great differences for some methods. In fact, approaches that assign contradictory classes to different messages sharing the same multimedia content are penalized in our second evaluation framework (drop in recall). In contrast, our LK−I2 approach benefits from its default strategy for media content classified as *unknown* by LK−I. The results of each approach are discussed in the following subsections.

B. Comparison of the different modalities

In addition to the quantitative results provided above, we examine the approaches according to the type of descriptors

that they exploit (text, source or image modality) and their possible complementarity for the fusion experiments presented in the following section. We exclude from this study predictions that already involve fusion between modalities. Thus only predictions LK−T, LK−I and LK−S will be kept among our predictions, MML−T, MML−I, MCG−T and MCG−I for the teams *MMLAB* and *MCG-ICT*. The predictions of the *VMU* team are all based on fusion (see Sec. IV-D). However, we retain VMU−S1 which is mainly based on the sources and which achieves the best performances. These eight predictions, hereafter denoted *elementary*, will be used in the following.

1) Text-based runs: Three predictions can be associated with a textual approach: LK−T, MML−T and MCG−T. The prediction LK−T tends to classify all messages as *false*, which can be explained by the strong imbalance of classes in the training set (three times more *false* than *true* messages). Thus, *636* real messages are classified as *false*. Conversely, the MML−T and MCG−T predictions tend to wrongly classify *false* messages as *true* (*i.e.* respectively 557 and 457 *false* messages on 1,230 are predicted *true*).

We can also note a difference between these three predictions depending on the descriptors used. While the predictions MML−T and MCG−T are based on shallow descriptors (essentially the set of descriptors $\mathcal{T}$), LK−T uses content descriptors (*i.e.* precise patterns). These predictions are thus possibly adapted to a fusion in order to merge their capacities.

2) Source-based runs: Two predictions are identified as using sources: LK−S and VMU−S1. While both approaches are based on a list of trusted sources, VMU−S1 also considers a source of non-trust. It can be noted that since the two lists of trusted sources are not identical, they can be complementary. A second difference is the choice of the assigned class in case of absence of source. While VMU−S1 chooses the *false* class, which is the majority class of the training set, LK−S makes the choice of the *unknown* class which inevitably gives a misclassified message (since 'no message actually has this class) but which allows a high accuracy of messages classified as *true* or *false* (respectively 100.00% and 92.97%) at the expense of the recall (respectively 41.22% and 87.47%).

3) Multimedia-based runss: Multimedia approaches are the most diverse. There are three predictions in which images or videos are used: LK−I, MML−I and MCG−I. Thus, even if the multimedia approaches have the weakest results individually, they may be complementarity.

LK−I predicts the *true* or *false* class only for a few messages (170 out of $2,228$), but obtains a high accuracy (97.30% on the class *false*). Messages for which no similar image was found get the *unknown* class. All messages with a video as illustration also receive the *unknown* class. MCG−I is the only approach dealing with video processing while messages with video content represent 48.43% of the dataset.

Several phenomena can explain the poor performance of these systems. First, in the case of a slight difference between the original image (*real*) and the modified image (*false*), the images are considered as similar by the search engine. This directly impacts LK−I and MML−I that both search for similar images in databases. Secondly, the images referenced on the specialized sites are sometimes altered: image stamped with texts like 'false', 'rumor' or 'true', or with drawing (*e.g.* a red circle on the photoshopped area to help the reader find it). The images broadcasted on the social networks also undergo this type of editing which lower the similarity between the query image and the database image.

From these results, it also seems that the use of a similar image search (LK−I and MML−I) provides more information than the use of descriptors $\mathcal{FOR}$ (MCG−I).

The poor results of the image-based LK−I2 approach can be explained in part by the small size of the image database. In fact, only about 25% of the images to be classified were present in the database at the time of submission of the results for the challenge. The large number of images for which no decision has been made (class *unknown*) strongly impacts the results in terms of recall.

VI. Fusion strategies

A. Simple fusion of the runs

A direct fusion of the predictions of Tab. II is first studied in this section. Each message is described by the predictions *true*, *false* or *unknown* of the different systems, in order to learn a combination of the predictions. The fusions of the predictions are then carried out by four classification methods:

1) linear SVM;
2) decision tree;
3) *Random Forest* (with 500 trees of depth 2);
4) neural network (a *Dropout* layer, a dense hidden layer of size 20 and an output layer with sigmoid activation function).

In addition to these classifiers, we indicate the results of a reference system corresponding to the majority vote on the predictions of the participants (*i.e.* among the predictions, the most frequently predicted class is associated with the message).

The evaluation protocol is a kind of *one-leave-out* but at the event level: Each classifier, is trained on all the messages of all the events except one event whose messages serve as a test set to evaluate the performances. We repeat the process, each time leaving out a different event, and average the results. The results are presented in Tab. III; asterisk denotes statistically significant results (Wilcoxon test with $p = 0.05$) compared to the reference system (majority vote).

It should be noted that the reference system does not allow to outperform the best predictions for the task, unlike classifiers using all the methods of the participants. This shows that not all predictions have the same importance and classifiers can learn appropriate weights for each method, or even more complex nonlinear combinations. As such, the best classifier (neural network) allows a significant increase in the rate of good classification, while offering more consistency (lower standard deviation of the performance measures).

Some messages are more difficult to classify than others. In Fig. 2, we report the distribution of messages according to the

Direct fusion	Majority	SVM	Decision tree	Random Forest	NN
F	82,6 (31,6)	90,9 (23,9)*	84.3 (28,8)	90,5(24,6)*	91,4(23,7)*
Acc.	84,0 (28,3)	95,1 (11,7)*	86,9 (23,0)*	95,1(12,9)*	96,3(10,5)*

TABLE III
AVG. F-SCORE AND ACCURACY (%, WITH STD-DEV) OF FUSIONS OF ELEMENTARY RUNS.

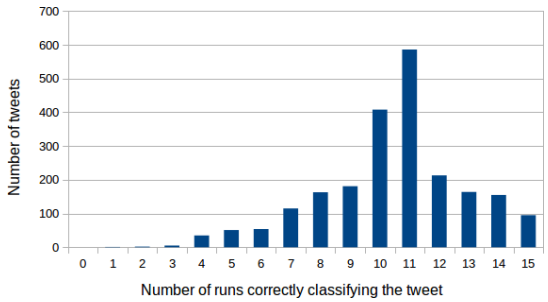

Fig. 2. Histogram of messages according to the number of runs correctly classifying them

number of methods classifying them correctly. All messages are correctly classified by at least one of the participants' methods. But some messages are incorrectly classified by most of the methods; In particular, there are 263 messages for which most methods are wrong. A simple fusion strategy will then have a great chance to rely on this majority to take its decision, which will lead to a prediction error. These difficult-to-categorize messages have one of three characteristics that may explain this difficulty:

1) messages written in non supported languages (information extraction failure) and making similarity calculations inappropriate (too few tweets in this language);
2) reduced URLs that hide the cited source (*e.g.* using short URL as goo.gl, t.co or bit.ly);
3) a large part of these messages come from events having both *true* and *false* messages and are therefore ambiguous (*Paris attacks* and *Fuji Lenticular*).

Two examples of such tweets are given in Fig. 3.

To study the contributions to the fusion of each of the methods, we can observe the produced classifiers. In the following, we focus on the *Random Forest* which gets both

Fig. 3. Example of tweets incorrectly classified by more than 12 methods of the VMU partcipants and by the the fusion

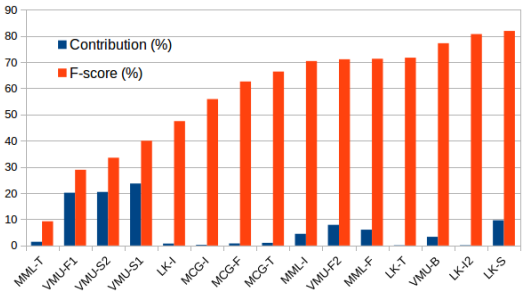

Fig. 4. Contribution (Gini index) of each system in the Random forest fusion, with the F-score of the system alone

good scores and allows to study these contributions easily. The contribution of an attribute (in our case the prediction of a method) is defined as the importance according to the Gini index, also called *mean decrease impurity* and as defined by [26], averaged over all trees in the random forest and normalized between 0 and 100%. We present these contributions in Fig. 4, and we compare them with the performances of the submissions taken independently.

It is worth noting that the best systems are not the ones preferentially used in the fusion. Indeed, `VMU-F1`, `VMU-S1` and `VMU-S2` contribute for more than 60% to the fusion, while their scores are among the lowest. These three systems are the most accurate, but have a low recall (many messages classified as *unknown*), which explains their low overall results. The fusion allows to exploit their very high precision when they predict *true* or *false*, and to refer to other systems otherwise.

We have seen that approaches can complement one another to improve prediction scores. However, the proposed fusion uses all the predictions while the information conveyed by each classifier can be redundant (*e.g.* predictions `MCG-T` and `MCG-I` affect the prediction `MCG-F`). Moreover, we do not obtain any information on the contributions of each modality during the direct fusion. We examine these two points in the following subsections.

B. Fusion elementary predictions

The results of a direct fusion of the eight elementary predictions defined previously (`LK-T`, `LK-I`, `LK-S`, `VMU-S1`, `MML-T`, `MML-I`, `MCG-T` and `MCG-I`; see Sec. V-B) are presented in Tab. IV. The reference system is again the majority vote on the eight input predictions. In the case of equality, the *unknown* class is used.

Despite the withdrawal of half of the input predictions, it is still possible to properly classify 95.0% of images and their associated messages. The fusion thus still brings an absolute gain

Fusion	Majority	SVM	Dec. Tree	Random Forest	NN
F	88.5	88.6	88.5	90.0*	91.3*
Acc.	93.3	92.8	92.3	95.0*	95.9*

TABLE IV
PERFORMANCE (%) OF FUSION ON THE 8 ELEMENTARY RUNS

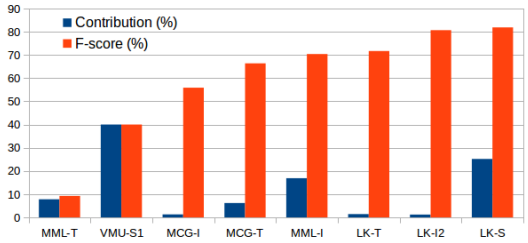

Fig. 5. Contribution (Gini index) of each system in the fusion by random forest, and their respective F-score.

of 10% compared to the best system (LK−S in this evaluation scenario). It is also interesting to compare these results with those in Tab. III. In particular, better results are obtained with the reference system by retaining only elementary predictions. This is easily explained since majority voting is sensitive to duplicates (and more broadly to correlations) induced by runs already including fusion. Classification methods that are not very sensitive to these correlation phenomena between attributes, such as *Random Forest*, logically obtain equivalent results. The fusion in this case is based on systems partly different from those seen previously, as can be seen in Fig. 5, but finally offers identical performance.

C. Importance of external knowledge

Some of the eight elementary systems use knowledge external to the training data. On the one hand, (LK−S and VMU−S1) are based on the identification of sources, which rely on white or black lists of sources that have been compiled manually. On the other hand, MML−I and LK−I approaches use external image databases for image search and comparison. It is thus legitimate to question the influence of this external knowledge in the results obtained, in particular because of the great contribution of source-based approaches in the previous experiment. We give in Tab. V the results obtained by the fusion restricted to elementary approaches using no external resource to the training data.

This time, the performance is lower than previous fusion attempts, and even lower than some of the methods taken separately (Tab. II). This last point shows that the four remaining methods predict different classes (this also indicated by the scores of the reference system), and that it is difficult to find a pattern to favor one method over another (scores of fusion by learning methods lower than the majority vote). Finally, the importance of the external resources used in some participants' systems is clear, since their absence results in a 25% drop in the performance of the fusion.

Fusion	Majority	SVM	Dec. tree	Random Forest	NN
F	63.3	60.0	59.4	60.4	61.1
Acc.	61.8	59.8	60.3	60.7	62.1

TABLE V
PERFORMANCES (%) OF FUSION FOR RUNS USING NO EXTERNAL KNOWLEDGE

1^{st} level prediction		SVM	Dec. tree	Random Forest	NN
Text	F	89.7	76.8	89.7	88.9
	Acc.	94.3	82.0	94.3	93.8
Source	F	89.6	83.2	89.6	89.2
	Acc.	93.9	87.9	93.9	93.9
Image	F	68.8	56.6	68.2	68.4
	Acc.	68.7	63.0	67.1	68.9

TABLE VI
PERFORMANCE (%) OF FUSION FOR DIFFERENT LEVELS AND MODALITIES

D. Fusion by modality

From the set of eight elementary runs, we propose a two-level fusion in which the first level classifies the messages according to the three modalities (text, source or image) and the second level regroups these three 1st-level predictions.

Tab. VI first presents the results of the three classifiers of first level (prediction at the level of the text, source or image). A first observation is the encouraging result of the classifier merging the text-based runs. Indeed, the results are significantly higher than those of individual systems. For the sources, the gain of the fusion is also present. Yet, the fusion of image-based approaches tends to produce worse results than the best image run.

To implement second-level fusion, we rely on neural networks, which are simple to implement and perform well in all previous fusion experiments. The architecture of the network reflects our two-level approach: the three neural networks, each corresponding to the fusion of approaches based on the same modality, serve to feed a network of second-level neurons (same architecture as in other experiences). To train this network, we test two approaches (noted training 1 and training 2 hereafter):

1) the text, image and source neural networks are trained individually, and the second-level network is then trained from their outputs;
2) the entire neural network is trained as a whole.

We present the results of the second fusion level with the two training strategies in Tab. VII. As can be seen, the results in both cases are very good, but there is an interest in training the entire network all at once rather than by level. The difference is statistically significant (Wilcoxon test with $p = 0.05$). With the training strategy 1, the results are comparable to those of a direct fusion of all the runs.

VII. CONCLUSION

In this article, we propose and examine several fusion strategies based on the predictions made by the four teams participating in the *Verifying Multimedia Use* task of the

2^{nd} level prediction	training 1	training 2
F-score	91.2	94.2*
Acc.	95.1	97.8*

TABLE VII
PERFORMANCE (%) OF NN-BASED TWO-LEVEL FUSION FOR THE TWO TRAINING STRATEGIES

Mediaeval 2016 evaluation campaign. Thus, we have seen that approaches based on the credibility of the source obtain good prediction scores but rely on external resources (white or black lists of sources) whose construction and maintenance may not seem credible in an application. on a very large scale (tweets from different countries, in different languages, for example). Approaches based on image analysis usually have disappointing individual results because of their inability to give a prediction on many cases. On the other hand, combined with other approaches, they may prove to provide complementary information improving the overall performance of a system. More generally, we have found that it is not necessarily the best individual scores approaches that contribute the most to the fusion system. The learning fusion systems we have proposed make it possible to exploit the high accuracy of certain systems while offsetting their weak recall with other methods. Finally, the main result of this article is the interest of proposing systems that combine different approaches with late fusion. The most powerful strategy seems to combine them by level, grouping first the methods working on the same type of information (text, image, source). An implementation of this two-level approach with a neural network gives indeed very good results, significantly better than the other approaches explored in this article.

Many research issues remain open after this work. We are currently developing datasets to compare existing approaches to more numerous and more various cases (tweets, but also articles of blogs or opinion sites and newspapers). These datasets are made available on the site http://hoaxdetector.irisa.fr/. From a technical point of view, future work will aim at correcting some problems, highlighted by our experiments, of systems based on images (*e.g.* modified images considered as similar to the original real image, images not found). We plan to expand the coverage of the image database. We are also exploring ways to improve the content comparison module to eliminate false positives, and to locate modified areas in these images [27]. From an application point of view, the presentation of information to the user must also be studied. It seems inappropriate that a system implements a strict censorship of messages deemed *false*, but the presentation of doubtful elements raises man-machine interface challenges, but also cognitive challenges (acceptance of the judgment of the machine), especially when the decision results from multiple systems combined by techniques that make it difficult to explain the final decision (especially for neural networks).

REFERENCES

[1] J. Golbeck and J. Hendler, "Inferring binary trust relationships in web-based social networks," *ACM Transactions on Internet Technology (TOIT)*, vol. 6, no. 4, pp. 497–529, 2006.

[2] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on twitter," in *2012 SIAM International Conference on Data Mining*, 2012.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," *Stanford InfoLab*, 1999.

[4] F. Mezaris, V. Patras, and M. Ioannis Bringay, "Online multi-task learning for semantic concept detection in video," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 186–190.

[5] H. Farid, *Photo Forensics*. The MIT Press, 2016.

[6] N. B. A. Warif, A. W. A. Wahab, M. Y. I. Idris, R. Ramli, R. Salleh, S. Shamshirband, and K.-K. R. Choo, "Copy-move forgery detection: Survey, challenges and future directions," *Journal of Network and Computer Applications*, vol. 75, pp. 259 – 278, 2016.

[7] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Large-scale evaluation of splicing localization algorithms for web images," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 4801–4834, 2017.

[8] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proceedings of the European Conference on Computer Vision*, 2016.

[9] S. Middleton, "Reveal project-trust and credibility analysis," *Mediaeval 2015 Workshop*, 2015.

[10] M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Detecting image splicing in the wild (web)," in *Multimedia & Expo Workshops (ICMEW)*, 2015.

[11] S. Middleton, "Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video," *Mediaeval 2015 Workshop*, 2015.

[12] T. Gottron, J. Schmitz, and S. Middleton, "Focused exploration of geospatial context on linked open data," in *3rd International Conference on Intelligent Exploration of Semantic Data (ICIESD)*, 2014.

[13] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Information Processing & Management*, 2015.

[14] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, S. E. Middleton, K. Andreadou, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," in *MediaEval 2016 Workshop*, 2016.

[15] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Transactions on Information Forensics and Security*, 2012.

[16] W. Li, Y. Yuan, and N. Yu, "Passive detection of doctored jpeg image via block artifact grid extraction," *89th Signal Processing*, 2009.

[17] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 227–236, 2011.

[18] C. Pasquini, F. Pérez-González, and G. Boato, "A benford-fourier jpeg compression detector," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 5322–5326.

[19] S. E. Robertson, S. Walker, and M. Hancock-Beaulieu, "Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive," in *7th Text Retrieval Conference (TREC)*, 1998.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computing Research Repository (CRR)*, 2014.

[21] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *4th International Conference on Learning Representations (ICLR)*, 2016.

[22] C. Boididou, S. Middleton, S. Papadopoulos, D.-T. Dang-Nguyen, M. Riegler, G. Boato, A. Petlund, and Y. Kompatsiaris, "The VMU participation @ verifying multimedia use 2016," in *MediaEval 2016 Workshop*, Amsterdam, 2016.

[23] Q.-T. Phan, A. Budroni, C. Pasquini, and F. De Natale, "A hybrid approach for multimedia use verification," in *MediaEval 2016 Workshop*, 2016.

[24] J. Cao, Z. Jin, Y. Zhang, and Y. Zhang, "Mcg-ict at mediaeval 2016: Verifying tweets from both text and visual content," *MediaEval 2016 Workshop*, 2016.

[25] C. Silverman, *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage*, C. Silverman, Ed. The European Journalism Centre (EJC), 2014.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, ser. Statistics/Probability Series. Belmont, California, U.S.A.: Wadsworth Publishing Company, 1984.

[27] C. Maigrot, E. Kijak, R. Sicre, and V. Claveau, "Tampering detection and localization in images from social networks: A CBIR approach," in *Image Analysis and Processing - ICIAP 2017 - 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part I*, 2017, pp. 750–761. [Online]. Available: https://doi.org/10.1007/978-3-319-68560-1_67