**Noname manuscript No.**
(will be inserted by the editor)

# Verifying Information with Multimedia Content on Twitter

## A Comparative Study of Automated Approaches

**Christina Boididou · Stuart E. Middleton · Zhiwei Jin · Symeon Papadopoulos · Duc-Tien Dang-Nguyen · Giulia Boato · Yiannis Kompatsiaris**

Received: date / Accepted: date

**Abstract** An increasing amount of posts on social media are used for disseminating news information and are accompanied by multimedia content. Such content may often be misleading or be digitally manipulated. More often than not, such pieces of content reach the front pages of major news outlets, having a detrimental effect on their credibility. To avoid such effects, there is profound need for automated methods that can help debunk and verify online content in very short time. To this end, we present a comparative study of three such methods that are catered for Twitter, a major social media platform used for news sharing. Those include: a) a method that uses textual patterns to extract

C. Boididou
Information Technologies Institute, CERTH
E-mail: boididou@iti.gr

S. E. Middleton
University of Southampton IT Innovation Centre
E-mail: sem@it-innovation.soton.ac.uk

Z. Jin
University of Chinese Academy of Sciences
E-mail: jinzhiwei@ict.ac.cn

S. Papadopoulos
Information Technologies Institute, CERTH
E-mail: papadop@iti.gr

D. Dang-Nguyen
University of Trento; Dublin City University
E-mail: dangnguyen@disi.unitn.it; duc-tien.dang-nguyen@dcu.ie

G. Boato
University of Trento
E-mail: giulia.boato@unitn.it

Y. Kompatsiaris
Information Technologies Institute, CERTH
E-mail: ikom@iti.gr

claims about whether a tweet is fake or real and attribution statements about the source of the content; b) a method that exploits the information that same-topic tweets should be also similar in terms of credibility; and c) a method that uses a semi-supervised learning scheme that leverages the decisions of two independent credibility classifiers. We perform a comprehensive comparative evaluation of these approaches on datasets released by the Verifying Multimedia Use (VMU) task organized in the context of the 2015 and 2016 MediaEval benchmark. In addition to comparatively evaluating the three presented methods, we devise and evaluate a combined method based on their outputs, which outperforms all three of them. We discuss these findings and provide insights to guide future generations of verification tools for media professionals.

**Keywords** Fake Detection · Verification · Credibility · Veracity · Trust · Social Media · Twitter · Multimedia

## 1 Introduction

Recent years have seen a tremendous increase in the use of social media platforms such as Twitter and Facebook as a means of sharing news content and multimedia, and as a source and sensor of trends and events [1]. The simplicity of the sharing process has led to large volumes of news content propagating over social networks and reaching huge numbers of readers in very short time. Especially multimedia posts (images, videos) can very quickly reach huge audiences and become viral due to the fact that they are easily consumed.

Given the speed of the news spreading process and the competition of news outlets and individual news sources to publish first, the verification of information and content is often carried out in a superficial manner or even completely neglected. This leads to the appearance and spread of large amounts of fake media content. In particular, when a news event breaks (e.g., a natural disaster), and new information and media coverage is of primary importance, news professionals turn to social media to source potentially interesting and informative content. It is exactly this setting, when the risk of fake content becoming widely disseminated is the highest. By *fake*, we refer to any publication or post with multimedia content that does not represent accurately the event that it refers to. It may be reposted content falsely associated with a current event, digitally manipulated content, computer-generated imagery presented as real imagery or speculations regarding the association of persons with a current event. In a similar way, by *real* we refer to posts with content that rightly represent the event they claim to. There are also posts that, despite sharing fake multimedia content, explicitly report that the content is fake (e.g. to warn readers) or they refer to it with a sense of humour; those are excluded from our study.

An example of reposted content is a widely shared photo of two young children hugging each other (Fig. 1 (a)). The image was claimed to depict a brother who is protecting his little sister during the earthquake in Nepal (April 2015), but it was later reported that it was taken a decade ago in


Fig. 1: (a) Siblings' photo from Vietnam reposted as being from Nepal earthquake. (b) Fake media content spread during the Malaysian airlines breaking news story.

a province of Vietnam by a professional photographer. In some cases, the consequences of fake content reaching a very large part of the population can be quite severe. For example, fake images became popular on social media after the Malaysia Airlines passenger flight disappeared on 8th March (Fig. 1 (b) illustrates an example). During the investigation of the plane trace, false alarms that the plane was detected came up. Taking into account how sensitive the case was, the circulation of this content deeply affected the people directly involved in it, such as the families of the passengers, causing emotional distress. An extended analysis on rumour propagation during the London riots [32] concluded that rumours typically start with someone tweeting about an incident, which then gets re-tweeted and reposted in a number of variations. An interactive representation[1] of the riots' rumours across time shows the velocity with which fake information propagated and the fact that hours, even days are needed to debunk such false claims.

There are several challenges that journalists face in the process of assessing the veracity of user-generated content. Notably, "traditional" digital media verification techniques employed by journalists [34], e.g. looking into the Exif metadata[2] of content or getting in touch with the person that published it, are often not possible or very slow due to the characteristics of social media platforms. For instance, Twitter and Facebook remove the Exif metadata from posted images, and Twitter accounts in most cases provide no contact information (e.g., email, telephone number). Furthermore, conventional image forensics approaches are hardly applicable due to the image resizing and recompression operations that are automatically applied by these social media platforms to all uploaded content [38].

[1] `http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter`

[2] Exif metadata contain information about the date, time and location an image was taken, the model of the device, and copyright information, which can be very useful when assessing the credibility of multimedia content [34].

The above challenges highlight the need for novel tools that can help news professionals assess the credibility of online content. To this end, we present and compare three automated approaches to solve this problem on Twitter. This study is based on the Verifying Multimedia Use (VMU) task [4] that was organized as part of the 2015 and 2016 editions of the MediaEval initiative[3], with the goal to benchmark methods on the problem of automatically predicting whether a tweet that shares multimedia content is misleading (fake) or trustworthy (real). The presented approaches [7, 21, 26] are the ones that competed in this task and include a) a method that uses attribution in tandem with fake and genuine claim extraction, b) a method that verifies tweets by exploring inter-tweet information, and c) one that uses a semi-supervised learning scheme. This study conducts and presents a comprehensive comparison between them in a more extended experimental setting, and draws actionable insights regarding the strengths and weaknesses of each method. In particular, the main contributions of this article include the following:

- We describe a new benchmark dataset based on a revised and extended version of the MediaEval dataset. Our revised dataset has duplicate and near-duplicate tweets removed and its cross-event balance improved to reduce content bias towards more popular events. This dataset is publicly available for other researchers to use for benchmarking.
- We report a new experimental analysis using this new benchmark dataset following a leave-one-event-out cross-validation scheme. Each of our fake classification approaches is evaluated and its ability to predict the veracity of content analysed and contrasted.
- We present results derived from an ensemble of the three fake classification approaches. We analyze the advantages of the ensemble-based approach and show that it is more effective in classifying fake content than each individual method on its own. This result provides a benchmark for other researchers to compare against in the future.

## 2 Background

Classifying online content with respect to its credibility and veracity is a highly complex problem that has been studied in multiple settings and using a variety of approaches. Our work focuses on the problem of *single post verification*, i.e. classifying an individual content item as being *fake* or *real*. This is in contrast to the related problem of *rumour detection* [41], which considers that a piece of false information is spreading across social networks. Although rumour detection is a highly relevant research problem and several methodological aspects are common with those arising in our problem setting, the following discussion is mostly focusing on single post verification approaches.

Several of the previous studies in the area focused on the statistical analysis on social media with the goal of extracting features that can be used as robust

[3] `http://multimediaeval.org/`

*verification indicators* for classifying social media content (sec. 2.1). Another field of study concerns methods for assessing the credibility of the source (or user account), where a post of interest originates (sec. 2.2). A different research area concerns the development of image forensics approaches that can potentially provide valuable complementary signals to the verification process (sec. 2.3). We also present in sec. 2.4 a few systems that attempt to solve the problem by leveraging methods and results from the works of sec. 2.1-2.3. Finally, we provide a description of the Verifying Multimedia Use (VMU) task (sec. 2.5) that constitutes the basis for our experimental study.

### 2.1 Verification cues for social media content

Several previous studies focus on the automatic extraction of credibility cues from the content of social media posts, either by using Natural Language Processing from the posts' text or by extracting other features. For instance, Castillo et al. [11] presented a supervised learning method to assess content credibility in Twitter. They extract discussion topics that are categorized as news or chat by human annotators and they build a model to automatically determine which topics are newsworthy by assigning a credibility label to them. Martinez-Romo et al. [25] retrieve tweets associated with trending topics and use blacklists to detect spam URLs in them. Then, they introduce language models based on probability distributions over pieces of text and by adding content features, they apply several models to evaluate their approach, which achieves high accuracy in classifying tweets with spam content. O'Donovan et al. [29] performed an analysis of the utility of various features when predicting content credibility. First, they collected Twitter data derived from very different contexts and they defined a set of features including content-based features, user profile features, and others that focus on the dynamics of information flow. Then, by checking the distribution of each feature category across Twitter topics, they concluded that their usefulness can greatly vary with context, both in terms of the occurrence of a particular feature, and the manner in which it is used. The work in [15], which is very similar in terms of objective to the VMU task that we study in this paper, tries to distinguish between fake and real images shared on Twitter by using decision tree-based classification models on tweet text and Twitter account features. Using Hurricane Sandy as the evaluation dataset, they report a 97% detection accuracy.

### 2.2 Source and user credibility on social networks

Several approaches focus on the study of user behaviour on social networks as well as on the detection of spam accounts. Stringhini et al. [36] investigated techniques for automated identification of spam accounts on Twitter by detecting anomalous behaviour. They used six features (friend-follower ratio, URL ratio in messages, similarity of messages sent by a user, friend choice, messages sent, friend number) in a classification model and managed to identify

about 15,000 spam accounts on Twitter. Canini et al. [9] proposed a method, that, given a particular topic, identifies relevant users, based on a combination of their expertise and trust. The authors employed an LDA topic model, using keywords extracted from the user profile, to estimate the association between a user account and a topic and then rank them based on their credibility. Additionally, Starbird et al. [35] examined an automated mechanism for identifying Twitter accounts providing eyewitness testimony from the ground. They used profile features, such as number of statuses and of followers, and features that describe how the Twitter community interacts with the user during the event. Finally, they applied an SVM classifier with asymmetric soft margins and they managed to achieve promising results on the task. Two of the methods proposed in this work also use features derived from social media accounts (including newly proposed features) with the aim of finding common characteristics of users that tend to share misleading content.

### 2.3 Image forensics

Image forensics has been long used for assessing the authenticity of images by detecting whether a digital image has been manipulated. Image manipulation is typically classified as splicing (transferring an object from an image and injecting it into another), copy-move (copying an object from the same image to a different position) or retouching (enhancing contrast, sharpening edges or applying color filters). These manipulations normally leave digital traces that forensics methods try to detect. The method in [14] exploits inconsistencies in the Color Filter Array (CFA) interpolation patterns, allowing for accurate splice localization. Since the most common format of digital images is JPEG, numerous methods try to exploit traces left by the JPEG compression process. In [24] and [31], different methods are proposed to determine whether an image was previously JPEG compressed. In [3], original and forged regions are discriminated in double compressed images for both aligned (A-DJPG) and non-aligned JPG (NA-DJPG). Double and triple compressions are detected and discriminated in [30], by exploiting the analysis of the Benford-Fourier coefficients. A method for tampered regions detection was proposed in [23] on the block artifact grids (BAG), which are caused by the block-based processing during JPEG compression and are usually mismatched in copy-move or splicing manipulations. Other methods aim to detect non-native JPEG images by analyzing quantization tables, thumbnails and information embedded in Exif metadata [22]. For copy-move detection, many methods have been proposed in the recent years, mainly based on the matching of keypoints [2], or regions [18]. For detecting image retouching, most current methods exploit illumination or shadow inconsistencies [28], or geometric relations disagreement [12] within an image. Despite the proliferation of image forensics methods, a recent experimental study [39] has concluded that many of them are ineffective on real cases of manipulated images sourced from the Web and social media due to the fact that such images typically undergo multiple resaving operations

that destroy a considerable part of the forensic traces. Nonetheless, one of the compared methods in our study makes use of forensics features (extracted from the image accompanying the tweet) as additional verification signals.

### 2.4 Systems for assessing content credibility

In the context of assessing content credibility, Ratkiewicz et al. developed the Truthy system [33] for real-time tracking of political memes on Twitter and for detecting misinformation, focusing on political astroturf. Truthy collects tweets, detects memes and introduces a web interface that lets users annotate the memes they consider truthful. Another system for evaluating Twitter content is TweetCred [16], a system that computes for each tweet a credibility score. It takes the form of a Web application that can be installed as a Chrome extension. The system encourages users to give feedback by declaring whether they agree or not with the produced score. We include TweetCred in our comparative experimental study as a state-of-the-art method.

### 2.5 Verifying Multimedia Use (VMU) task

To assess the effectiveness of automated tweet verification methods, we rely on resources produced by the VMU task [4], which was introduced in 2015 as part of the MediaEval benchmarking initiative. The definition of the task is the following: *"Given a tweet and the accompanying multimedia item (image or video) from an event of potential interest for the international news audience, return a binary decision representing verification of whether the multimedia item reflects the reality of the event in the way purported by the tweet."* In practice, participants received a list of tweets that include images or video and were required to automatically predict, for each tweet, whether it is trustworthy or deceptive (`real` or `fake` respectively). An `unknown` label is also accepted in case that there is no available prediction for a tweet. In addition to fully automated approaches, the task also considered human-assisted approaches provided that they are practical (i.e., fast enough) in real-world settings, such as manually identifying the veracity of a multimedia item by searching on trustworthy online websites or resources. The following considerations should be made in addition to the above definition:

- A tweet is considered **fake** when it shares multimedia content that does not faithfully represent the event it refers to. The variety of untrustworthy and misused content appearing in the context of past events led us to devise a small typology of misleading use of multimedia content (see Fig. 2).
- A tweet is considered to be **real** when it shares multimedia that accurately represents the event it refers to.
- A tweet that shares content that does not represent accurately the event it refers to but reports the false information or refers to it with a sense of humour is **neither considered fake nor real** (and hence not included in the datasets released by the task).

Fig. 2: Different types of misleading multimedia use. From left to right: a) reposting an old photo showing soldiers guarding the Tomb of the Unknown Soldier claiming it was captured during the Hurricane Sandy in 2012, b) reposting digital artwork as a photo from the solar eclipse in March 2015, c) speculation of depicted people as being suspects of the Boston Marathon bombings in 2013, d) spliced sharks on a photo captured during the Hurricane Sandy.

For each tweet, the task has also released three types of feature:

- tweet-based (`TB-base`): Extracted from the tweet itself, e.g. the number of terms, the number of mentions and hashtags, etc. [8].
- user-based (`UB-base`): Based on the Twitter profile, e.g. the number of friends and followers, the account age, whether the user is verified, etc. [8].
- forensics (`FOR`): Forensic features extracted from the visual content of the tweet image, and specifically the probability map of the aligned double JPEG compression, the potential primary quantization steps for the first six DCT coefficients of the non-aligned JPEG compression, and the PRNU (Photo-Response Non-Uniformity) [13].

## 3 Description of verification approaches

### 3.1 Using attribution, fake and genuine claim extraction (UoS-ITI)

This approach is motivated by an established journalistic process for verifying social media content [34]. The central hypothesis is that the "wisdom of the crowds" is not really wisdom when it comes to verifying suspicious content. Instead it is better to rank evidence from Twitter according to the most trusted and credible sources in a way similar to the one practiced by journalists.

A trust and credibility model was created based on an NLP pipeline involving tokenization, Part-Of-Speech (POS) tagging, Named Entity Recognition (NER) and Relation Extraction (RE). The novelty of this approach lies within the choice of regex patterns, which are modelled on how journalists verify fake and genuine claims by looking at the source attribution for each claim, and the semi-automated workflow allowing trusted lists of entities to be utilized. A novel conflict resolution approach was created based on ranking claims in order of trustworthiness. To extract fake and genuine claims, a set of regex patterns were created (see Fig. 3) matching both terms and POS tags. Claims of an image being fake or genuine occur infrequently, and by themselves are

not sufficient. If an image is claimed to be real without any supporting attribution we assume it is fake, since from our own analysis strong debunking posts almost always contain attribution. We combine all fake and real claims with trusted source attribution (e.g. via BBC News) to discover strong evidence. To extract attribution, a combination of Named Entity (NE) matching, based on noun and proper noun sequences, and regex patterns for source citation was used. Other researchers have published linguistic patterns that were used to detect rumours [8,10,40], but the combination of fake/genuine claims and source attribution used by the UoS-ITI approach is novel in that it uses insights from well-established journalistic processes for social media content.

First, an NLP pipeline is executed (see Fig. 4) that takes in each tweet from the test dataset and tokenizes it using a Punkt sentence tokenizer and a Treebank word tokenizer. To help the POS tagger, no stemming is applied and text case is preserved. Each tokenized sentence is passed to a Treebank POS tagger, which supports more European multi-lingual tagsets than other taggers (e.g., Stanford POS tagger), an important consideration for future

| **Named Entity Patterns** | **Examples** |
|---|---|
| @ <ANY> | @bbcnews, BBC News, CNN.com, CNN |
| (NOUN \| PROP_NOUN \| NAMESPACE) | |
| (NOUN \| PROP_NOUN \| NAMESPACE) (NOUN \| PROP_NOUN \| NAMESPACE) | |
| **Attribution Patterns** | |
| ...<NE> <SYMBOL> <URl> | What a great picture! @bbcnews: http://bit.\|y/1234 |
| <NE> *{0,1} <IMAGE> *(0,2) <URl> ... | @bbcnews image - http://bit.\|y/1234 |
| <NE> *{0,1} <FROM> *{0,2} <URl> ... | @bbcnews releases photo http://bit.\|y/1234 |
| <RT> <SYMBOL>{0,1} <NE> ... | RT: @bbcnews "\|ove\|y picture of ... |
| ... <FROM> *{0,2} <NE> | ... eyewitness report via @bbcnews |
| **Faked Patterns** | |
| ... <IMAGE> *{0,1} ^<POSnot> <FAKED> ... | ... image is fake! ... |
| ... <POSis> <POSa>{0,1}^<POSnot> <FAKED> *{0,1} <IMAGE> ... | ... is a fake image ... |
| ... <POSis>{0,1} <POSnot> <POSa>{0,1} <REAL> ... | ... Is not a real ... |
| **Genuine Patterns** | |
| ... <IMAGE> <POSis> *{0,1} ^<POSnot> <REAL> ... | ... image is totally genuine ... |
| ... <POSis> <POSa>{0,1} ^<POSnot> <REAL> *{0,1} <IMAGE> ... | ... is a real image ... |
| ... <POSis>{0,1} <POSnot> <POSa>{0,1} <FAKE> ... | ... Is not a fake ... |

**Key**

| | |
|---|---|
| (mm) = n to m matches allowed | * = any non-whitespace characters |
| <NE> = named entity | ^ = forbit match |
| <SYMBOL> = symbols (e.g. : = -) | <IMAGE> = image variants (e.g. image, video) |
| <POSnot> = POS adverbs RB (e.g. not, never) | <FROM> = from variants (e.g. via, attributed) |
| <POSis> = POS verbs VBZ, VBD (e.g. is, was) | <FAKED> = fake variants (e.g. fake, hoax) |
| <POSa> = POS determiner DT (e.g. a, the) | <REAL> = real variants (e.g. real, genuine) |
| | <RT> = RT variants (e.g. RT, MT) |

Fig. 3: Verification Linguistic Patterns in UoS-ITI. These patterns are encoded as regex patterns matching on both phrases in content and their associated POS tags (e.g. NN = noun, NNP = proper noun).

work as breaking news can happen anywhere in the world, not just English speaking locations. Namespaces and URI's are extracted prior to tokenization and re-inserted after POS tagging as explicit NEs so they can be matched using regex expressions later.

The employed NER strategy is based on a regex POS expression that matches unigrams and bigrams with nouns, proper nouns, namespaces and Twitter usernames. This is a high recall-low precision approach to NER as we want to capture all relevant NEs at this stage. Next, a lookup table is used to filter candidate NEs into sets of trusted, untrusted and unknown entities; this allows the removal of known false positive NE values and the application of blaklist and whitelist values. Finally, candidate NEs are used to create the POS and NE-labelled sentence, which is passed to a set of regex expressions encoding typical relationship phrases for fake, real and attribution claims.

The approach is semi-automated in that it exploits a list of a priori known trusted and untrusted sources. All news providers have long lists of trusted sources for different regions around the world so this information is readily available. For this task, a list of candidate NEs was created by first running the NER regex patterns on the test dataset. Then, each NE was manually checked via Google search (e.g. looking at Twitter profile pages) and NEs were removed that were considered as irrelevant for inclusion in a list of trusted or untrusted sources by a journalist. Instead, NEs were kept that included news organizations, respected journalists and well cited bloggers and experts. Creating these lists took under two hours (570 NEs checked, 60 accepted).

The employed RE approach uses a set of regex expressions that match serialized POS and NE tagged sentence trees. These regex expressions were manually created after a detailed analysis of the linguistic patterns from Twitter, YouTube and Instagram around a number of previously crawled event types (e.g., hurricanes, tornados, earthquakes, blackouts, conflicts, etc.). For finding attributed NEs, all the common ways were studied, in which social media users

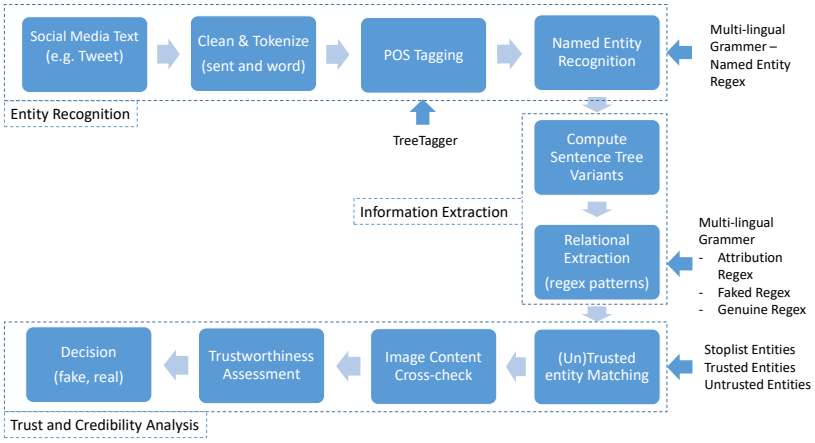


Fig. 4: NLP pipeline for regex-based NER and RE in UoS-ITI.

attribute sources. This is typically either as a reference to a NE followed by a link or image reference, or a statement claiming that a link or image reference is created and/or verified by an NE. For fake and genuine claims, regex patterns are created from the most commonly phrased claims about images or links being fake or real. Examples of the regex patterns can be seen in Fig. 3 and make heavy use of the POS tags to avoid being overly term specific.

Once all the attributed sources, fake and genuine claims have been identified, a trustworthiness decision for each image is made. Just like human journalists, claims are ranked by trustworthiness based on whether the claim comes directly from a trusted author (top rank), is attributed to a trusted source (second rank) or is from an unknown source (third rank). Claims directly by, or attributed to, untrusted sources are ignored. The final decision for each image is taken based on only the most trustworthy claims. A conservative claim conflict resolution approach is used, where a fake claim by a source supersedes a real claim by an equally trusted source.

## 3.2 Using a two-level classification model (MCG-ICT)

Existing approaches often formulate the tweet verification problem as a binary classification task [8]. Features from tweet text 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 the real world have strong relations among each other, especially, tweets of the same topic will likely have the same credibility: real or fake.

Rather than classifying each tweet individually, the MCG-ICT approach verifies tweets by leveraging inter-tweet information, such as whether they contain the same multimedia content. It was empirically observed that even such simple implications among tweets would be useful to boost the original message-level predictions. In fact, in recent work [19,20], links among tweets were built by clustering tweets into sub-events or topics. Thus, credibility evaluation can be performed at different scales to provide more robust predictions.

### *3.2.1 Two-level classification model*

As illustrated in Fig. 5, the MCG-ICT approach comprises two levels of classification: a) The message-level, which learns a credibility model per message (tweet). Features extracted from the text content, user information and other components of a tweet are used for training a classifier. b) The topic-level, i.e. a specific subject of discussion or sub-event under the broader unfolding event. By assuming tweets under a same topic likely have similar credibility values, tweets are clustered into different topics. Compared with raw tweets, topics eliminate variations of tweets by aggregating message-level credibility classifications. The topic-level feature is computed as the average of the tweet-level feature vectors around the topic. The following processing steps take place for topic-level classification:

- *Topic clustering*: In [19], a clustering algorithm is used to cluster tweets into sub-events. But this algorithm performs poorly in forming topics in the target dataset as it is difficult to decide the optimal number of clusters. However, in the studied verification setting, each tweet contains an image or video, and each image/video can be contained in more than one tweets. This intrinsic one-to-many relation is used to form topics: each image/video corresponds to a topic and tweets containing it are assigned to this topic.
- *Topic labeling*: Each topic is labelled using the majority of the labels of its tweets. These labels are used for training the topic-level classifier. In fact, with the proposed topic formation method, almost all tweets in a topic have the same label, resulting in topics labelled with very high confidence.
- *Topic-level feature aggregation*: Message-level features (section 3.2.2) of all tweets in a topic are aggregated by averaging them to derive a single topic-level feature vector. By taking the average of all tweets, the impact of noisy or outlier tweets is suppressed.
- *Fusing topic-level result*: After topic-level classification, a probability value is computed for each topic representing the likelihood of it being fake. Then, for each tweet in the topic, this value is added as a feature to its original feature vector. Finally, a message-level classifier is trained with this extended feature in order to produce the final results.

In terms of classification model, several options from the state of the art were tested, and the selection was based on the performance on the development set using cross validation. J48 Decision Trees were selected for the topic-level classification, and Random Forests for the message-level classification.

### *3.2.2 Feature extraction*

At the message level, we use as base features the ones shared by the task, `TB-base` and `UB-base` (Sec. 2.5). Some additional features were also tested

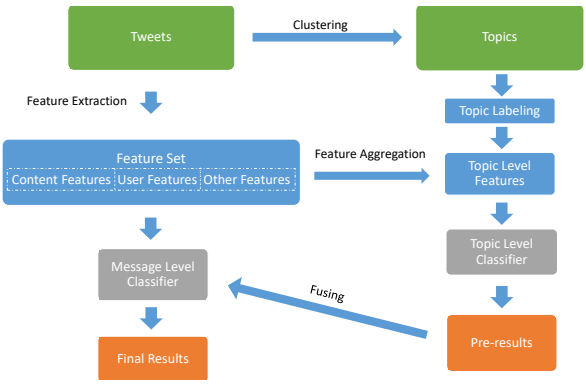


Fig. 5: Overview of the MCG-ICT two-level classification model. Topic-level classifications are fused with the message-level ones to produce the final result.

but not included in the final approach configuration: word term features and several image features.

The commonly used term frequency (tf) and tf-idf features were tested. Experiments on the training (development) set indicated that such features could lead to overfitting, since they led to very high performance on general cross-validation and very low performance on leave-one-event-out cross-validation. Since few words co-occur across different events, one may assume that other keyword-based features (e.g., LDA) would also contribute little to this task.

Several image-based features (e.g. image popularity, resolution) were also tested. Such image features could 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 showed that these features led to slightly worse performance for the topic-level classification, compared to content-based ones, and to much worse performance when combined with message-level features. Moreover, image features cannot be applied directly on videos included in the test set. Hence, those were not further considered.

### 3.3 Using an agreement-based retraining scheme (CERTH-UNITN)

This approach combines different sets of tweet-based (`TB`), user-based (`UB`) and forensics (`FOR`) features in a semi-supervised learning scheme. A more detailed exposition of this method is presented in [5]. The approach builds on supervised classification models and an agreement-retraining method that uses part of its own predictions as new training samples with the goal of adapting to tweets posted in the context of new events.

Fig. 6 depicts an overview of the method. It relies on two individual classification models, one based on the combination of `TB` and `FOR` features and a second based on `UB` features. Bagging is used to ensure higher reliability in the training process of the classifiers ($CL_{11}$ ... $CL_{1n}$ and $CL_{21}$ ... $CL_{2n}$), and an agreement-based retraining strategy (fusion) is employed with the goal of improving the accuracy of the overall framework. All classifiers are based on Random Forests of 100 trees.

#### *3.3.1 Feature extraction*

The approach uses the `TB-ext` and `UB-ext` features, which are an extended version of the `TB-base` and `UB-base` released by the MediaEval task. For the `FOR` features, we also include additional ones.

`TB-ext`: These are binary features extracted from the tweet text, e.g. the presence of a word, symbol or external link. Language-specific binary features are also used corresponding to the presence of specific terms; for languages, in which such terms are not available, the values of these features are set to null (missing). Language detection is performed with a publicly available library[4],

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

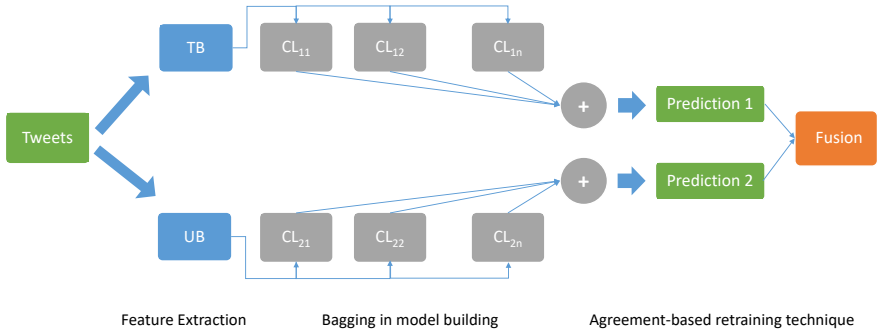


Fig. 6: Overview of the CERTH-UNITN method.

and a feature is added for the `number of slang words` in a text, using slang lists in English[5] and Spanish[6]. For the `number of nouns`, the Stanford parser[7] is used to assign POS tags to each word (only in English). For the readability of text, the Flesch Reading Ease method is used[8], which computes the complexity of a piece of text as a score in $[0, 100]$ (0: hard-to-read, 100: easy-to-read).
`UB-ext`: User-specific features are extracted such as `number of media content`, `account age` and others that refer to information that the profile shares. In addition, these include whether the user declares a location and whether this can be matched to a city name from the Geonames dataset[9].

For both `TB` and `UB` features, trust-oriented features are computed for the links shared, through the tweet itself (`TB`) or the user profile (`UB`). These include the WOT metric[10], a score indicating how trustworthy a website is according to reputation ratings by Web users, the in-degree and harmonic centrality, which are rankings based on the links of the web forming a graph[11], and web metrics provided by the Alexa API[12].
`FOR`: For each image, 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 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 (max, min, mean, median, most frequent value, st. deviation, and variance) are computed from the `BAG` values and concatenated to a 14-d vector. Figure 7 illustrates the feature extraction process. We apply the same process on the `NAJPG` map to obtain a second feature vector.

[5] `http://onlineslangdictionary.com/word-list/0-a/`
[6] `http://www.languagerealm.com/spanish/spanishslang.php`
[7] `http://nlp.stanford.edu/software/lex-parser.shtml`
[8] `http://simple.wikipedia.org/wiki/Flesch_Reading_Ease`
[9] `http://download.geonames.org/export/dump/cities1000.zip`
[10] `https://www.mywot.com/`
[11] `http://wwwranking.webdatacommons.org/more.html`
[12] `http://data.alexa.com/data?cli=10&dat=snbamz&url=google.gr`

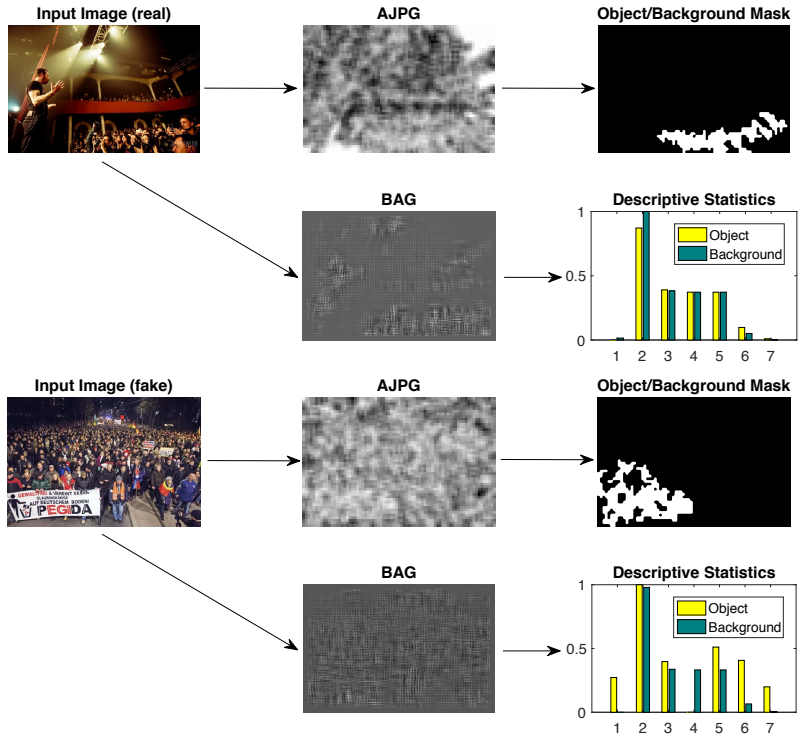


Fig. 7: Illustration of forensics feature extraction process.

### *3.3.2 Data pre-processing and bagging*

To handle the issue of missing values on the features, Linear Regression (LR) is used for interpolating missing values. This is applied only to numeric features as the method is not applicable for boolean values. Only feature values from the training set are used in this process. Data normalization is performed to scale values to the range [-1, 1]. Furthermore, bagging is used to improve the accuracy of the method. 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 ($m = 9$), as shown in Fig. 6. The final prediction for each of the test samples is calculated using the majority vote of the $m$ predictions.

### *3.3.3 Agreement-based retraining*

*Agreement-based retraining* is used to improve the prediction accuracy for unseen events. This is motivated by a similar approach implemented in [37] on the problem of polarity classification. To this end, two classifiers are built $CL_1$, $CL_2$, each on different types of feature, and their outputs are combined

as follows: The two outputs for each sample of the test set are compared, and depending on their agreement, the test set is divided in two subsets, the *agreed* and *disagreed* sets. Assuming that the agreed predictions are correct with high likelihood, they are used as training samples to build a new classifier for classifying the disagreed set of instances. To this end, in the subsequent step, the agreed samples are added to the best performing of the two initial models, $CL_1$, $CL_2$ (comparing them on the basis of their performance using cross-validation on the training set). The goal of this method is to retrain the initial model and adapt it to the 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.

## 4 Experiments and Evaluation

### 4.1 Datasets

The conducted experiments were based on the benchmark dataset released by the VMU task in 2015 (Sec. 2.5). We refer to the original version of the dataset as `dataset#1`. This has been collected over a number of years using a crowd-sourcing approach. Images are found by volunteers (including the authors of the article), and paid micro-workers (via Amazon Mechanical Turk). Each suggested content item was provided with associated news reports or debunking articles by journalists, offering evidence regarding its veracity. These were then used to provide ground truth labels (`real`/`fake`) for each tweet.

The dataset consists of tweets relating to 17 events listed in Table 1, comprising in total 197 cases of real and 191 cases of misused images, associated with 6,225 real and 9,404 fake tweets posted by 5,895 and 9,025 unique Twitter users respectively. Note that several of the events, e.g., Columbian Chemicals, Passport Hoax and Rock Elephant, were actually hoaxes, hence all multimedia content associated with them was fake. For several real events (e.g., MA flight 370) no real images (and hence no `real` tweets) are included in the dataset, since none came up as a result of the conducted data collection process.

For further testing, we additionally created `dataset#2`, which is a subset of `dataset#1` by first performing near-duplicate tweet removal: we empirically set a minimum threshold of similarity and computed the Levenshtein Distance[13] for each pair of texts. A small amount of near-duplicate texts exceeding the threshold were manually removed. Note that in `dataset#1` the number of unique fake and real multimedia items, which the tweets are associated with, is highly unbalanced. As the aim of `dataset#2` is to create a balanced dataset, we randomly selected a subset of fake and real multimedia items as well as the tweets associated with them.

Finally, to assess the stability of results, we also compared the methods on the dataset that was used in the VMU task sequel in 2016 [6]. This is a superset of `dataset#1`, using the latter as development set, while it contains

[13] `http://rosettacode.org/wiki/Levenshtein_distance#Java`

an additional 998 real and 1,230 fake tweets in the test set, organized around 64 cases of real and 66 cases of misused multimedia items. The tweet IDs and image URLs for all of the above datasets are publicly available[14].

Table 1: **Upper part:** MediaEval'15 events and derivative datasets (`#1`, `#2`): For each event, we report the numbers of unique real (if available) and fake images ($I_R$, $I_F$ respectively), unique tweets that shared those images ($T_R$, $T_F$) and unique Twitter accounts that posted those tweets ($U_R$, $U_F$). **Bottom part:** MediaEval'16 events and corresponding statistics.

| ID | Event | dataset#1 | | | | | | dataset#2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $I_R$ | $T_R$ | $U_R$ | $I_F$ | $T_F$ | $U_F$ | $I_R$ | $T_R$ | $U_R$ | $I_F$ | $T_F$ | $U_F$ |
| E1 | Hurricane Sandy | 150 | 4,664 | 4,446 | 53 | 5,558 | 5,432 | 60 | 838 | 825 | 16 | 376 | 369 |
| E2 | Boston Marathon bombing | 29 | 344 | 310 | 35 | 189 | 187 | 18 | 131 | 120 | 13 | 56 | 56 |
| E3 | Sochi Olympics | - | - | - | 14 | 274 | 252 | - | - | - | 9 | 76 | 74 |
| E4 | MA flight 370 | - | - | - | 23 | 310 | 302 | - | - | - | 13 | 88 | 87 |
| E5 | Bring Back Our Girls | - | - | - | 7 | 131 | 126 | - | - | - | 6 | 35 | 33 |
| E6 | Columbian Chemicals | - | - | - | 15 | 185 | 87 | - | - | - | 6 | 124 | 64 |
| E7 | Passport hoax | - | - | - | 2 | 44 | 44 | - | - | - | 1 | 5 | 5 |
| E8 | Rock Elephant | - | - | - | 1 | 13 | 13 | - | - | - | 1 | 4 | 4 |
| E9 | Underwater bedroom | - | - | - | 3 | 113 | 112 | - | - | - | 2 | 4 | 4 |
| E10 | Livr mobile app | - | - | - | 4 | 9 | 9 | - | - | - | 3 | 6 | 6 |
| E11 | Pig fish | - | - | - | 1 | 14 | 14 | - | - | - | 1 | 4 | 4 |
| E12 | Solar Eclipse | 5 | 140 | 133 | 6 | 137 | 135 | 2 | 55 | 54 | 3 | 53 | 53 |
| E13 | Girl with Samurai boots | - | - | - | 3 | 218 | 212 | - | - | - | 2 | 16 | 16 |
| E14 | Nepal Earthquake | 11 | 1004 | 934 | 20 | 356 | 343 | 6 | 113 | 107 | 8 | 178 | 176 |
| E15 | Garissa Attack | 2 | 73 | 72 | 2 | 6 | 6 | 2 | 40 | 39 | 2 | 4 | 4 |
| E16 | Syrian boy | - | - | - | 1 | 1786 | 1692 | - | - | - | 1 | 197 | 195 |
| E17 | Varoufakis and zdf | - | - | - | 1 | 61 | 59 | - | - | - | 1 | 29 | 28 |
| | Total | 197 | 6,225 | 5,895 | 191 | 9,404 | 9,025 | 88 | 1,177 | 1,145 | 88 | 1,255 | 1,178 |

| Event | $I_F$ | $T_F$ | $I_R$ | $T_R$ |
|---|---|---|---|---|
| Gandhi Dancing | 1 | 29 | - | - |
| Half of Everything | 9 | 39 | - | - |
| Hubble Telescope | 1 | 18 | - | - |
| Immigrants fear | 5 | 33 | 3 | 18 |
| ISIS children | 2 | 3 | - | - |
| John Guevara | 1 | 33 | - | - |
| Mc Donalds Fee | 1 | 6 | - | - |
| Nazi Submarine | 2 | 11 | - | - |
| North Korea | 2 | 10 | - | - |
| Not Afraid | 2 | 32 | 3 | 35 |
| Pakistan Explosion | 1 | 53 | - | - |
| Pope Francis | 1 | 29 | - | - |
| Protest | 1 | 30 | 10 | 34 |
| Refugees | 4 | 35 | 13 | 33 |
| Rio Moon | 1 | 33 | - | - |
| Snowboard Girl | 2 | 14 | - | - |
| Soldier Stealing | 1 | 1 | - | - |
| Syrian Children | 1 | 12 | 1 | 200 |
| Ukrainian Nazi | 1 | 1 | - | - |

| Event | $I_F$ | $T_F$ | $I_R$ | $T_R$ |
|---|---|---|---|---|
| Woman 14 children | 2 | 11 | - | - |
| American Soldier Quran | 1 | 17 | - | - |
| Airstrikes | 1 | 24 | - | - |
| Attacks in Paris | 3 | 44 | 22 | 536 |
| Ankara Explosions | - | - | 3 | 19 |
| Bush book | 1 | 27 | - | - |
| Black Lion | 1 | 7 | - | - |
| Boko Haram | 1 | 31 | - | - |
| Bowie David | 2 | 24 | 4 | 48 |
| Brussels Car Metro | 3 | 41 | - | - |
| Brussels Explosions | 3 | 69 | 1 | 9 |
| Burst in KFC | 1 | 25 | - | - |
| Convoy Explosion Turkey | - | - | 3 | 13 |
| Donald Trump Attacker | 1 | 25 | - | - |
| Eagle Kid | 1 | 334 | - | - |
| Five Headed Snake | 5 | 6 | - | - |
| Fuji Lenticular Clouds | 1 | 123 | 1 | 53 |
| Total | 66 | 1,230 | 64 | 998 |

## 4.2 Measuring accuracy by leave-one-event-out cross-validation

The conducted experiments aimed at evaluating the accuracy of each method on a variety of unseen events. The features of fake tweets may vary across different events, so the generalization ability of automated methods is considered an important aspect of its verification performance. To this end, we used

[14] `https://github.com/MKLab-ITI/image-verification-corpus/`

each time one of the events $E_i$, $i = \{1, 2, ..., 17\}$ for testing, and the remaining ones for training. For example, for evaluating the performance on event $E_1$, we used the tweets of $E_2$, $E_3$, ..., $E_{17}$ for training, and the tweets of $E_1$ for testing. This is in contrast to the MediaEval task, where events E1-11 were used for training, and E12-17 for testing. To evaluate the approaches, we used the established measures of *precision* ($p$), *recall*, and *F1*-score ($F1$). Assuming that the *positive* class is the case that a tweet instance is `fake`, and *negative* that a tweet instance is `real`, we define these metrics as:

$$p = \frac{tp}{tp + fp}, \quad F1 = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{1}$$

where $tp$ refers to true positives (correctly detected `fake`), $fp$ to false positives (`real` misclassified as `fake`), and $fn$ to false negatives (`fake` as `real`).

Table 2 presents the precision and $F1$-scores that the approaches achieved in the context of the MediaEval tasks and on `dataset#1` and `dataset#2`. We also compare our results with those presented by the TweetCred method [16]. We have re-implemented a variation of the method described in the paper: we identified the common tweet- and user-based features that the CERTH-UNITN and TweetCred methods use, and we built classification models for each of the datasets. We evaluated common state-of-the-art classifiers on each dataset: SVM, RandomForest and AdaBoost. For the first two datasets (`MediaEval '15` and `dataset#1`), SVM achieved the highest performance, while for the rest RandomForest worked best. As can be seen, TweedCred ranks second on two of the datasets (`dataset#2` and `MediaEval'16`), third on `MediaEval'15` and first on `dataset#1`.

Table 2: Precision and $F1$-scores achieved on a) MediaEval VMU 2015 task, with E1-11 used for training and E12-17 for testing, b) `dataset#1`, c) `dataset#2`, and d) MediaEval VMU 2016 task. For b and c, the leave-one-event-out cross-validation method was used for measuring performance.

| **Method** | `MediaEval '15` | | `dataset#1` | | `dataset#2` | | `MediaEval '16` | |
|---|---|---|---|---|---|---|---|---|
| | *p* | *F1* | *p* | *F1* | *p* | *F1* | *p* | *F1* |
| **UoS-ITI** | **1.000** | 0.830 | **0.938** | 0.224 | **0.917** | 0.244 | 0.520 | 0.468 |
| **MCG-ICT** | 0.964 | **0.942** | 0.804 | 0.756 | 0.816 | **0.750** | 0.563 | 0.504 |
| **CERTH-UNITN** | 0.861 | 0.911 | 0.755 | 0.693 | 0.690 | 0.635 | **0.980** | **0.911** |
| **TweedCred [16]** | 0.680 | 0.800 | 0.810 | **0.820** | 0.810 | 0.650 | 0.580 | 0.720 |

Fig. 8 illustrates the $F1$-scores of the tested approaches for each event of `dataset#1` and `dataset#2`. The mean performance of the approaches is also illustrated in red. Events are ordered based on the $F1$-score achieved by the highest-scoring method (MCG-ITI). On average, it is clear that on `dataset#1` the two-level classification method (MCG-ITI) outperforms the other two. However, given that it strongly relies on the fact that tweets are correlated when they share the same multimedia content, it seems that it performs better on events that comprise only one or few unique multimedia cases (e.g. Passport


Fig. 8: $F1$-scores for `dataset#1` and `dataset#2` per event and approach.

hoax, Syrian boy). When a single event includes numerous multimedia cases (e.g., Nepal earthquake), its performance decreases. The UoS-ITI approach seems to accurately predict the posts associated with video content (Syrian boy, Varoufakis and zdf). Similar results are obtained on `dataset#2` (Fig. 8).

In addition to $F1$-scores, we also report the corresponding precision scores in Fig. 9 for each event on `dataset#1` and `dataset#2`. In both cases, the UoS-ITI approach outperforms the other two, providing a very high precision ($> 0.9$) for the small set of tweets it was able to classify. This is an important observation, since it allows us to use the results of UoS-ITI as a type of pre-classifier in an ensemble-based approach (Sec. 4.4).

Another striking finding concerns the stability of performance of the three methods when testing them on a different dataset (Mediaeval '16). The CERTH-UNITN approach is a clear winner in this case, since it is the only approach that manages to retain its performance at comparable levels, while the other two approaches perform considerably worse. This provides evidence in support of the generalization ability of the agreement-based retraining method.

### 4.3 Measuring verification performance per multimedia item

A further experiment explores the verification accuracy of methods on each unique multimedia item. Given that `dataset#1` contains 388 unique multimedia items, we divided tweets in groups according to the multimedia item they are associated with. Then, we calculate the performance of each approach on each of those unique multimedia cases. Fig. 10 illustrates the achieved $F1$-scores. In the horizontal axis, we present the unique multimedia items. The


Fig. 9: Precision scores for `dataset#1` and `dataset#2` per event and approach.

figure reveals that the MCG-ITI and CERTH-UNITN approaches perform similarly in the majority of cases, while the UoS-ITI achieves quite low performance compared to them, due to the fact that it avoids producing a result in cases where there is not sufficient information to make this decision.

Another key question of this analysis is the level of overlap between the method results. For this reason, we conduct pairwise comparisons between the previously generated $F1$-score distributions. After calculating the number of multimedia items for which the methods have an $F1$-score equal to zero and equal to one, we report the percentage of items, for which the methods' predictions agree. In addition, we calculate the Pearson correlation between these distributions. Table 3 presents the results of this analysis. These make clear that the three approaches are highly correlated in terms of the cases they fail to predict ($F1$-score $= 0$) and less correlated in the cases where they succeed ($F1$-score $= 1$). The pairwise Pearson correlations demonstrate that MCG-ICT and CERTH-UNITN approaches are much more similar in their predictions compared to UoS-ITI. Overall, these experiments reveal that there is potential for improving upon the results of the individual methods by fusing their results, which we investigate in the next section.

### 4.4 An ensemble verification approach

We investigate three ensemble methods for fusing the results of the three approaches:

– **ENS-MAJ**: For each tweet, we aggregate individual predictions by majority vote. However, in the UoS-ITI approach, several of the predictions

Table 3: Pairwise comparison of $F1$-score distributions. Reporting percentages (%) where the $F1$-score of both methods is equal to 0 and 1, and the Pearson correlation between them.

| | *F1=0* | *F1=1* | *pearson corr* |
|---|---|---|---|
| UoS-ITI vs MCG-ICT | 53.3 | 0.7 | 0.295 |
| UoS-ITI vs CERTH-UNITN | 59.5 | 6.7 | 0.332 |
| MCG-ICT vs CERTH-UNITN | 51.5 | 9.5 | 0.707 |

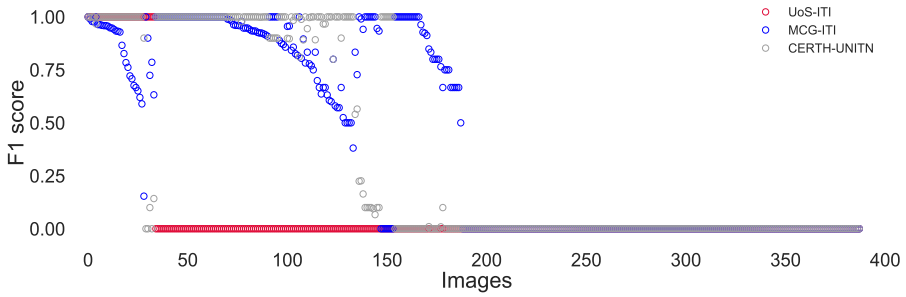


Fig. 10: $F1$-score across unique images for each approach. To make the visualization of results cleaner, images in the $x$ axis are sorted based on the mean F-score across the three approaches.

are marked as `unknown`, which is a problem in case the decisions of the other two approaches disagree. To overcome this tie, we assign the tweet as *fake*, assuming that it is preferable to falsely consider a case to be `fake` than falsely consider it as `real`.

- **ENS-HPF**: This takes advantage of the high precision (low fp rate) of the UoS-ITI method. If the UoS-ITI prediction for a tweet is other than `unknown`, we adopt it as correct; otherwise, we check the other methods' predictions and in case of disagreement we consider the item to be `fake`.
- **ENS-ORA**: This is a hypothetical (oracle) ensemble method, which selects the correct prediction if at least one of the methods' predictions is correct. This provides an upper-bound of the performance that could be theoretically possible if we could optimally combine the three approaches.

Fig. 11 presents the resulting $F1$-scores and precision per event of the ensemble methods. On average, these achieve higher performance than the individual approaches. This result stems from the complementarity of the approaches' outputs, which was illustrated in sec. 4.3, and the effectiveness of the proposed schemes in combining their outputs. **ENS-ORA** delineates the maximum possible performance that is achievable by combining the three methods. Out of the two practical fusion schemes, **ENS-HPF** produces more accurate predictions compared to **ENS-MAJ** as the former uses the highly accurate UoS-ITI approach as the preferred approach for performing the classification and falls back to the other two approaches in case UoS-ITI produces no result.


Fig. 11: $F1$-scores and Precision of the `ENS-MAJ`, `ENS-HPF` and `ENS-ORA` ensemble methods for `dataset#1`, mean score of each method on the dataset of MediaEval '16 and individual mean scores of the approaches.

### 4.5 Relevance of experimental results for real-world use cases

These promising results from our ensemble verification approach should also be seen in the context of typical use cases for automated verification of images and videos trending on social media.

An example use case involving semi-automated verification is in support of journalists who are trying to verify social media content for use in news stories. Breaking news in particular has competing objectives to publish content first (i.e. as quickly as possible) and get the verification right (i.e. take enough

time to ensure that the content is not fake). Publishing eyewitness content before rivals will gain a journalist much kudos. Publishing a fake image, then later being forced to retract the story, can seriously damage a journalist's reputation. Let us consider a use case where a journalist wants to receive a live feed (e.g. every 5 or 10 minutes) of the top 500 trending images on Twitter, classified and filtered using our ensemble of fake classifiers. Our best reported result ($F1$=0.79, $p$=0.82, $r$=0.81) means that on average for 500 trending images, only 90 would be classified in error. During news events such as the Hurricane Sandy 2012 fake social media images on Twitter [27] [17] outnumbered real images by two to one. In the context of our use case this means that of the 500 images considered, 333 would on average be fake, and of those 333, 273 would be classified as fake and filtered. This represents a significant reduction in the images the journalist needs to consider in real-time, something which is very useful when working under breaking news deadlines where a story must be verified and published within minutes.

Another example user case, this time involving fully automated verification, is where an automated news summarization platform wants to aggregate news feeds from sources such as popular social media bloggers in real-time. Readers of such news summarization platforms typically accept a higher number of false stories than they would from a journalist-based news service. In this case the volume of trending images that need checking would be much larger, with tens of thousands of images being checked as candidates for aggregation into news alert summary snippets. The fully automated nature of our approach makes this a viable proposition. Even for platforms like Storyful, where news snippets are passed to a human checkdesk for a final verification step, our approach could have significant utility as a pre-filter.

## 5 Conclusions

In this article, we presented a comparative study for automated approaches for verifying online information with multimedia content. By presenting three different in nature methods, we showed that there are several ways to deal with the challenging problem of verification. To measure verification accuracy, we evaluated these methods by using leave-one-out cross-validation and by reporting their scores per multimedia case. In the `MediaEval'15` dataset and `dataset#2`, the MCG-ICT method achieved the highest F1-scores, particularly for events with few cases of unique multimedia items. The UoS-ITI achieved the highest precision scores with a very low false positive rate for the events it could classify. The CERTH-UNITN method led to consistently high results on average, and in particular on events with many tweets. Importantly, it managed to retain its performance on the `MediaEval'16` dataset, clearly outperforming the other two and the TweetCred method, and demonstrating the potential of the agreement-based retraining scheme for making the approach applicable to new datasets.

By combining approaches into an ensemble method we were able to further increase the F1-score of the best performing method by approximately 7% (**ENS-HPF**) with a theoretical upper bound of improvement of approximately 20% (**ENS-ORA**), which is a very encouraging finding, and provides a state of the art benchmark for other researchers in this field. To improve results even further we feel that more sophisticated use of visual and contextual features are needed. This is a very challenging area and will need a combination of image forensics, computer vision and cross-referencing of contextual information (e.g. weather, maps, etc.) about the event of interest.

**Acknowledgements** This work has been supported by the REVEAL and InVID projects, partially funded by the European Commission (FP7-610928 and H2020-687786 respectively).

## References

1. Aiello, L.M., Petkos, G., Martín, C.J., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. IEEE Trans. Multimedia **15**(6), 1268–1282 (2013). DOI 10.1109/TMM.2013.2265080. URL http://dx.doi.org/10.1109/TMM.2013.2265080
2. Ardizzone, E., Bruno, A., Mazzola, G.: Copy move forgery detection by matching triangles of keypoints. Information Forensics and Security, IEEE Transactions on **10**(10), 2084–2094 (2015)
3. Bianchi, T., Piva, A.: Image forgery localization via block-grained analysis of jpeg artifacts. IEEE Transactions on Information Forensics and Security **7**(3), 1003–1017 (2012)
4. Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y.: Verifying multimedia use at mediaeval 2015. In: MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany (2015)
5. Boididou, C., Papadopoulos, S., Apostolidis, L., Kompatsiaris, Y.: Learning to detect misleading content on twitter. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 278–286. ACM (2017)
6. Boididou, C., Papadopoulos, S., Dang-Nguyen, D., Boato, G., Riegler, M., Middleton, S.E., Petlund, A., Kompatsiaris, Y.: Verifying multimedia use at mediaeval 2016. In: Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016. (2016). URL http://ceur-ws.org/Vol-1739/MediaEval_2016_paper_3.pdf
7. Boididou, C., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Kompatsiaris, Y.: The certh-unitn participation@ verifying multimedia use 2015 (2015)
8. Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., Newman, N.: Challenges of computational verification in social multimedia. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 743–748 (2014). DOI 10.1145/2567948.2579323. URL http://dx.doi.org/10.1145/2567948.2579323
9. Canini, K.R., Suh, B., Pirolli, P.L.: Finding credible information sources in social networks based on content and social structure. In: Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom), pp. 1–8. IEEE (2011)
10. Carton, S., Adar, E., Park, S., Mei, Q., Zeffer, N., Resnick, P.: Audience analysis for competing memes in social media. In: Ninth International AAAI Conference on Web and Social Media (2015)
11. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 675–684. ACM (2011)
12. Conotter, V., Boato, G., Farid, H.: Detecting photo manipulation on signs and billboards. In: Image Processing, IEEE International Conference on, pp. 1741–1744 (2010)

13. Conotter, V., Dang-Nguyen, D.T., Riegler, M., Boato, G., Larson, M.: A crowdsourced data set of edited images online. In: Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, CrowdMM '14, pp. 49–52. ACM, New York, NY, USA (2014). DOI 10.1145/2660114.2660120. URL http://doi.acm.org/10.1145/2660114.2660120
14. Ferrara, P., Bianchi, T., Rosa, A.D., Piva, A.: Image forgery localization via fine-grained analysis of cfa artifacts. IEEE Transactions on Information Forensics and Security **7**(5), 1566 –1577 (2012)
15. Gupta, A., Kumaraguru, P.: Twitter explodes with activity in mumbai blasts! a lifeline or an unmonitored daemon in the lurking? (2012)
16. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: Tweetcred: A real-time web-based system for assessing credibility of content on twitter. In: Proc. 6th International Conference on Social Informatics (SocInfo) (2014)
17. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web companion, pp. 729–736. International World Wide Web Conferences Steering Committee (2013)
18. Jian, L., Xiaolong, L., Bin, Y., Xingming, S., Li, J., Li, X., Yang, B., Sun, X.: Segmentation-Based Image Copy-Move Forgery Detection Scheme. Information Forensics and Security, IEEE Transactions on **10**(3), 507–518 (2015)
19. Jin, Z., Cao, J., Jiang, Y.G., Zhang, Y.: News credibility evaluation on microblog with a hierarchical propagation model. In: 2014 IEEE International Conference on Data Mining (ICDM), pp. 230–239. IEEE (2014)
20. Jin, Z., Cao, J., Zhang, Y., Luo, J.: News verification by exploiting conflicting social viewpoints in microblogs. In: AAAI 2016. AAAI (2016)
21. Jin, Z., Cao, J., Zhang, Y., Zhang, Y.: Mcg-ict at mediaeval 2015: Verifying multimedia use with a two-level classification model (2015)
22. Kee, E., Johnson, M.K., Farid, H.: Digital image authentication from jpeg headers. IEEE Transactions on Information Forensics and Security **6**(3-2), 1066–1075 (2011)
23. Li, W., Yuan, Y., Yu, N.: Passive detection of doctored jpeg image via block artifact grid extraction. IEEE Transactions on Signal Processing **89**(9), 1821–1829 (2009)
24. Luo, W., Huang, J., Qiu, G.: JPEG error analysis and its applications to digital image forensics. IEEE Transactions on Information Forensics and Security **5**(3), 480 –491 (2010)
25. Martinez-Romo, J., Araujo, L.: Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications **40**(8), 2992–3000 (2013)
26. Middleton, S.E.: Extracting attributed verification and debunking reports from social media: Mediaeval-2015 trust and credibility analysis of image and video (2015)
27. Middleton, S.E., Middleton, L., Modafferi, S.: Real-time crisis mapping of natural disasters using social media. Intelligent Systems, IEEE **29**(2), 9–17 (2014)
28. O'Brien, J.F., Farid, H.: Exposing photo manipulation with inconsistent reflections. ACM Transactions on Graphics **31**(1), 4:1–4:11 (2012)
29. O'Donovan, J., Kang, B., Meyer, G., Hollerer, T., Adalii, S.: Credibility in context: An analysis of feature distributions in twitter. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 293–301. IEEE (2012)
30. Pasquini, C., Boato, G., Perez-Gonzalez, F.: Multiple jpeg compression detection by means of benford-fourier coefficients. In: Proceedings of the Workshop on Information Forensics and Security. IEEE (2014)
31. Pasquini, C., Perez-Gonzalez, F., Boato, G.: A benford-fourier jpeg compression detector. In: Proceedings of the International Conference on Image Processing, pp. 5322–5326. IEEE (2014)
32. Procter, R., Vis, F., Voss, A.: Reading the riots on twitter: methodological innovation for the analysis of big data. International Journal of Social Research Methodology **16**(3), 197–214 (2013)
33. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of

the 20th international conference companion on World wide web, pp. 249–252. ACM (2011)
34. Silverman, C.: Verification handbook: a definitive guide to verifying digital content for emergency coverage (2013)
35. Starbird, K., Muzny, G., Palen, L.: Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: Proc. 9th Int. Conf. Inf. Syst. Crisis Response Manag. Iscram (2012)
36. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9. ACM (2010)
37. Tsakalidis, A., Papadopoulos, S., Kompatsiaris, I.: An ensemble model for cross-domain polarity classification on twitter. In: Web Information Systems Engineering–WISE 2014, pp. 168–177. Springer (2014)
38. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Detecting image splicing in the wild (WEB). In: 2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015, pp. 1–6 (2015). DOI 10.1109/ICMEW.2015.7169839
39. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications pp. 1–34 (2016)
40. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: Early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1395–1405. International World Wide Web Conferences Steering Committee (2015)
41. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: A survey. arXiv preprint arXiv:1704.00656 (2017)