MONITOR: A Multimodal Fusion Framework to Assess Message Veracity in Social Networks

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jérôme Darmont, Camille Noûs

MONITOR: A Multimodal Fusion Framework to Assess Message Veracity in Social Networks

Abderrazek Azri[1], Cécile Favre[1], Nouria Harbi[1], Jérôme Darmont[1], and Camille Noûs[2]

[1] Université de Lyon, Lyon 2, UR ERIC
5 avenue Pierre Mendès France, F69676 Bron Cedex, France
[2] Université de Lyon, Lyon 2, Laboratoire Cogitamus
{a.azri, cecile.favre, nouria.harbi, jerome.darmont}@univ-lyon2.fr
camille.nous@cogitamus.fr

Abstract. Users of social networks tend to post and share content with little restraint. Hence, rumors and fake news can quickly spread on a huge scale. This may pose a threat to the credibility of social media and can cause serious consequences in real life. Therefore, the task of rumor detection and verification has become extremely important. Assessing the veracity of a social media message (e.g., by fact checkers) involves analyzing the text of the message, its context and any multimedia attachment. This is a very time-consuming task that can be much helped by machine learning. In the literature, most message veracity verification methods only exploit textual contents and metadata. Very few take both textual and visual contents, and more particularly images, into account. In this paper, we second the hypothesis that exploiting all of the components of a social media post enhances the accuracy of veracity detection. To further the state of the art, we first propose using a set of advanced image features that are inspired from the field of image quality assessment, which effectively contributes to rumor detection. These metrics are good indicators for the detection of fake images, even for those generated by advanced techniques like generative adversarial networks (GANs). Then, we introduce the Multimodal fusiON framework to assess message veracIty in social neTwORks (MONITOR), which exploits all message features (i.e., text, social context, and image features) by supervised machine learning. Such algorithms provide interpretability and explainability in the decisions taken, which we believe is particularly important in the context of rumor verification. Experimental results show that MONITOR can detect rumors with an accuracy of 96% and 89% on the MediaEval benchmark and the FakeNewsNet dataset, respectively. These results are significantly better than those of state-of-the-art machine learning baselines.

Keywords: Social networks · Rumor verification · Image features · Machine learning.

1 Introduction

After more than two decades of existence, social media has attracted a large number of users. These social platforms allow users to share content and interact with each other. They enable the rapid diffusion of information in real-time, regardless of its credibility, for two main reasons: first, there is a lack of a means to verify the veracity of the content transiting on social media; and second, users often publish messages without verifying the validity and reliability of the information. Consequently, social networks, and particularly microblogging platforms, are a fertile ground for rumors to spread.

(a) Black clouds in New York City before Sandy!!!

(b) #NepalEarthquake 4Years old boy protect his little sister. make me feel so sad

Fig. 1: Two sample rumors posted on Twitter

Following previous work [1], we define a rumor as an item of circulating information whose veracity status is yet to be verified at posting time. Widespread rumors can pose a threat to the credibility of social media and cause harmful consequences in real life. Thus, the automatic assessment of information credibility on microblogs that we focus on is crucial to provide decision support to, e.g., fact checkers. This task requires to verify the truthfulness of messages related to a particular event and return a binary decision stating whether the message is true. In the literature, most automatic rumor detection approaches address the task as a classification problem. They extract features from various aspects of messages, which are then used to train a wide range of machine learning [30] or deep learning [31] methods. Features are generally extracted from the textual content of messages [24] and the social context [33]. However, the multimedia content of messages, particularly images that present a significant set of features, are little exploited.

In this paper, we second the hypothesis that the use of image properties is important in rumor verification. Images may indeed attract more attention than texts [2]. Furthermore, images play a crucial role in the news diffusion process. For example, in the dataset collected by [12], the average number of messages with an attached image is more than 11 times that of plain text ones. Figure 1 shows two sample rumors posted on Twitter. In Figure 1(a), it is hard to assess veracity from the text, but the likely-manipulated image hints at a rumor. In Figure 1(b), it is hard to assess veracity from both the text or the image because the image has been taken out of its original context. Based on the above observations, we aim to leverage all the modalities of microblog messages for verifying rumors; that is, features extracted from textual and social context content of messages, and up to now unused visual and statistical features derived from images. Then, all types of features must be fused to allow a supervised machine learning classifier to evaluate the credibility of messages.

Our contribution is twofold. First, we propose the use of a set of image features inspired from the field of image quality assessment (IQA) and we prove that they contribute very effectively to the verification of message veracity. These metrics estimate the rate of noise and quantify the amount of visual degradation of any type in an image. They are proven to be good indicators for detecting fake images, even those generated by advanced techniques such as generative adversarial networks (GANs) [9]. To the best of our knowledge, we are the first to systematically exploit this type of image features to check the veracity of microblog posts. Our second contribution is the Multimodal fusiON framework to assess message veracIty in social neTwORks (MONITOR), which exploits all types of message features (i.e., text, social context and image features) by supervised machine learning. This choice is motivated by two factors. First, these techniques provide explainability and interpretability about the decisions taken. We believe that such explanations are necessary, especially in the context

of rumors, with people's privacy in line. Second, we do also want to explore the performance of deep machine learning methods in the near future, especially to study the tradeoff between classification accuracy, computing complexity, and explainability.

Eventually, extensive experiments conducted on two real-world datasets demonstrate the effectiveness of our rumor detection approach. MONITOR indeed outperforms all state-of-the-art machine learning baselines with an accuracy and F1-score of up to 96% and 89% on the MediaEval benchmark [6] and the FakeNewsNet dataset [26], respectively.

The rest of this paper is organized as follows. In Section 2, we first review and discuss related works. In Section 3, we detail MONITOR and especially feature extraction and selection. In Section 4, we present and comment on the experimental results that we achieve with respect to state-of-the-art methods. Finally, in Section 5, we conclude this paper and outline future research.

2 Related Works

2.1 Non-image Features

Studies in the literature present a wide range of non-image features. These features may be divided into two subcategories, textual features and social context features. Textual features are extracted from the text content of messages, they are derived from the linguistics of a text to capture specific writing styles and the headlines that commonly occur in fake news content, such as lexical and syntactic features.

To classify a message as fake or real, Castillo *et al.* [8] capture prominent statistics in tweets, such as count of words, capitalized characters and punctuation, total number of words and characters. Beyond these features, lexical words expressing specific semantics or sentiments are also crucial clues to characterize the text, emotional marks (question marks and exclamation marks), and emoticons are also counted. Many sentimental lexical features are proposed in [16], who utilize a sentiment tool called the Linguistic Inquiry and Word Count (LIWC) to count words in meaningful categories.

Other works exploit syntactic features derived from the sentence level of rumors, such as the number of keywords, the sentiment score or polarity of the sentence. Features based on topic models are used to understand messages and their underlying relations within a corpus. Wu *et al.* [32] train a Latent Dirichlet Allocation model [5] with a defined set of topic features to summarize semantics for detecting rumors on the Sina Weibo microblogging platform.

The social context reflects the interactions among different users and describes the propagating process of a rumor [27]. Post content features represent the users' social response in terms of stance. Social network features are extracted by constructing specific networks, such as diffusion [16] or co-occurrence networks [25].

Recent approaches detect fake news based on temporal-structure features. Kwon *et al.* [15] studied the stability of features over time and found that, for rumor detection, linguistic and user features are suitable for early-stage, while structural and temporal features tend to have good performance in the long-term stage.

2.2 Image Features

Although images are widely shared on social networks, their potential for verifying the veracity of messages in microblogs is not sufficiently explored. Morris *et al.* [22] assume that the user profile image has an important impact on information credibility published by this user. For images

attached in messages, very basic features are proposed by [32], who define a feature called "has multimedia" to mark whether the tweet has any picture, video or audio attached. Gupta *et al.* [10] propose a classification model to identify fake images on Twitter during Hurricane Sandy. However, their work is still based on textual content features.

To automatically predict whether a tweet that shares multimedia content is fake or real, Boididou *et al.* [6] propose the Verifying Multimedia Use (VMU) task. Textual and image forensics [17] features are used as baseline features for this task. They conclude that Twitter media content is not amenable to image forensics and that forensics features do not lead to consistent VMU improvement [7]. Finally, Jin *et al.* [2] mostly focus on classification models for the problem rather than image features.

3 MONITOR

Microblog messages contain rich multimodal resources, such as text contents, surrounding social context, and attached image. Our focus is to leverage this multimodal information to determine whether a message is true or false. Based on this idea, we propose a framework for verifying the veracity of messages. MONITOR's detailed description is presented in this section.

3.1 Multimodal Fusion Overview

We define a message as a tuple of text, social context, and image content. MONITOR takes features from these modalities and aims to learn a multimodal fusion features vector as an aggregation of these aspects of the message. Figure. 2 shows a general overview of MONITOR.

It has two main stages: 1) Features extraction and selection. We extract several useful features from the message text and the social context, we then perform a feature selection algorithm to identify the relevant features, which form a first set of textual features. From the attached image, we drive statistics and efficient visual features inspired from the IQA field, which form a second set of image features; 2) Model learning. Textual and image features sets are then concatenated and normalized to form the fusion vector as the final multimodal representation of the message. Several machine learning classifiers may learn from the fusion vector to distinguish the veracity of the message (i.e., real or fake).

3.2 Feature Extraction and Selection

The feature extraction stage aims to represent message content and related auxiliary information in a formal measurable structure. To better choose features, we reviewed the best practices followed by information professionals (e.g., journalists) in verifying content generated by social network users. We based our thinking on relevant data from journalistic studies [19] and the verification handbook [28]. We define a set of features that are important to extract discriminating characteristics of rumors. These features are mainly derived from three principal aspects of news information: content, social context, and visual content of images.

As for the feature selection process, it will only be applied to content and social context features sets to remove the irrelevant features that can negatively impact performance. Because our focus is the visual features set, we keep all these features in the learning process.

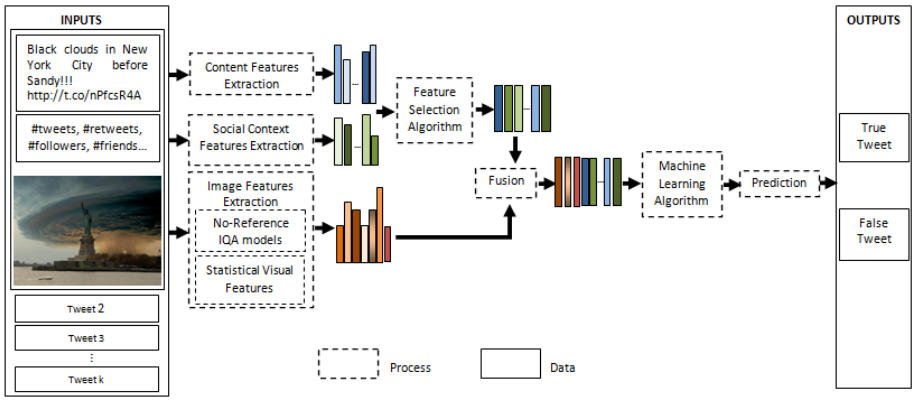


Fig. 2: Overview of MONITOR

Message Content Features Content features are extracted from the message's text. Aiming to arouse much attention and stimulate the public mood, rumor texts tend to have certain patterns in contrast to non-rumors. We extract characteristics such as the length of a tweet text and the number of its words. These characteristics also include statistics such as the number of exclamation and question marks, as well as binary features indicating the existence or not of emoticons. Furthermore, other features are extracted from the linguistics of a text, including the number of positive and negative sentiment words. For the English language, we use Liu and Hu's opinion lexicon list[3], for German the Leipzig Affective Norms [3], and for Spanish the adaptation of ANEW [4]. Additional binary features indicate whether the text contains personal pronouns.

The veracity of the message text could also be related to its readability. We calculate a readability score between 1 and 100 using the Flesch Reading Ease method [14], the higher this score is, the easier the text is to read. For tweets written in a language for which the above features cannot be extracted, we consider the corresponding values to be missing. Other features are extracted from the informative content provided by the specific communication style of the Twitter platform, such as the number of retweets, mentions(@), hashtags(#), and URLs.

Social Context Features The social context reflects the relationship between the different users and describes the process of spreading a rumor, therefore the characteristics of the social context are extracted from the behavior of the users and the propagation network. We capture several features from the users' profiles, such as number of followers and friends, number of tweets the user has authored, the number of tweets the user has liked, whether the user is verified by the social media, and whether the user has a profile image.

We extract, also, features from the propagation tree that can be built from tweets and re-tweets of a message, such as the depth of the re-tweet tree. Tables 1 and 2 depicts a description of a sets of content feature, and social context features extracted for each message.

To improve the performance of MONITOR, we perform a feature selection algorithm on the features sets listed in Tables 1 and 2. The details of the feature selection process are discussed in Section 4.

[3] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html#lexicon

Table 1: Content features

Description
of characters, words
of question mark (?), exclamation mark (!)
of uppercase characters in the tweet text
of positive, negative sentiment words
of mentions(@username), hashtags(#link), URLs
of happy, sad mood emoticon
first, second, third order pronoun
The readability score of the tweet text

Table 2: Social context features

Description
of followers, friends, posts the user has
Friends/followers ratio, times listed the user has
of re-tweets, likes that the tweet has obtained
Whether the user shares a homepage URL
Whether The user has their own profile image
Whether the author has a verified account
of Tweets the user has liked

Image Features To differentiate between false and real images in messages, we propose to exploit visual content features and visual statistical features that are extracted from the joined images.

Visual Content Features. Usually, a news consumer decides the image veracity based on his subjective perception, but how do we quantitatively represent the human perception of the quality of an image?. The quality of an image means the amount of visual degradations of all types present in an image, such as noise, blocking artifacts, blurring, fading, and so on.

The IQA field aims to quantify human perception of image quality by providing an objective score of image degradations based on computational models [18]. These degradations are introduced during different processing stages, such as image acquisition, process, compression, storage, transmission, decompression, display or even printing. Inspired by the potential relevance of IQA metrics for our context, we use these metrics in an original way for a purpose different from what they were created for. More precisely, we think that the quantitative evaluation of the quality of an image could be useful for veracity detection.

IQA is mainly divided into two areas of research: first, full-reference evaluation; and second, no-reference evaluation. Full-reference algorithms compare the input image against a pristine reference image with no distortion. In no-reference algorithms, the only input is the image whose quality we want to measure, these algorithms compare statistical features of the input image against a set of features derived from an image database.

In our case, we do not have the original version of the posted image; therefore, the approach that is fitting for our context is the no-reference IQA metric. For this purpose, we use three no-reference algorithms that have been demonstrated to be highly efficient.

The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [20] is trained on a database of images with known distortions, and is limited to evaluating the quality of images with the same type of distortion. BRISQUE is opinion-aware, which means that subjective quality scores accompany the training images.

The Naturalness Image Quality Evaluator (NIQE) [21] is trained on a database of pristine images and can measure the quality of images with arbitrary distortion. NIQE is opinion-unaware and does not use subjective quality scores.

The Perception based Image Quality Evaluator (PIQE) [29] is opinion-unaware and unsupervised (i.e., it does not require a trained model). PIQE can measure the quality of images with arbitrary distortion.

For example, Figure 3 displays the BRISQUE score computed for a natural image and its distorted versions (compression, noise and blurring distortions). The BRISQUE score is a non-

negative scalar in the range [1, 100]. Lower values of score reflect better perceptual quality of image.

Original image 13.7215 | JPEG compression 22.6603 | Gaussian Noise 28.5840 | Median Blur 4.1562

Fig. 3: BRISQUE score computed for a natural image and its distorted versions

No-reference IQA metrics are not only used for traditional forgery detection, but are also good indicators for other types of image modifications, such as GAN-generated images. These techniques allow modifying the context and semantics of images in a very realistic way. Unlike many image analysis tasks, where both reference and reconstructed images are available, images generated by GANs may not have any reference image. This is the main reason for using NR-IQA for evaluating this type of fake images, since these algorithms assess image quality without needing a reference nor its characteristics. Figure 4 displays the BRISQUE score computed for real and fake images generated by image-to-image translation based on GANs [34].

Real image 17.7778 | Fake image 22.0260 | Real image 12.5000 | Fake image 22.5279

Fig. 4: BRISQUE score computed for real and fake generated GAN images

Statistical Features. These are statistics of images attached to the text of a message. Similar to the statistical features of message content, some basic statistics of images proved to be distinctive in separating rumors and non-rumors. We define four statistical features from two aspects.

Number of Images: In addition to the text, a user can post one, several, or no images. To denote this feature, we count the total number of images in a rumor event and the ratio of posts containing more then one image.

Spreading of Images: During an event, some images are very replied and generate more comments than others. The ratio of such images is calculated to indicate this feature. Table 3 illustrates the

description of proposed visual and statistical features to characterize the content of images. We use the whole set of these features in the learning process.

3.3 Model Training

So far, we have obtained a first set of relevant textual features through a feature selection process performed on text content and social context features. We have also obtained a second set of image features composed of statistical and visual features driven from the field of image quality evaluation. These two sets of features are scaled, normalized, and concatenated to form the multimodal representation for a given message, which is fed to learn a supervised classifier for the rumor verification goal. Several learning algorithms can be implemented for the classification task of message veracity. In the experimental part, we investigate the algorithms that provide the best performance.

In summary, MONITOR takes as input training data with contents from three different modalities: text, social context and image. The output is the prediction label for each message instance to indicate it as true or false. In the following section, we present empirical experiments to evaluate MONITOR's ability to assess message veracity.

Table 3: Description of image features

Type	Feature	Description
Visual Features	BRISQUE	The BRISQUE score of a given image
	PIQE	The PIQE score of a given image
	NIQE	The NIQE score of a given image
Statistical Features	Count_Img	The number of all images in a news event
	Ratio_Img1	The ratio of the multi-image tweets in all tweets
	Ratio_Img2	The ratio of image number to tweet number
	Ratio_Img3	The ratio of the most widespread image in all distinct images

4 Experiments

In this section, we conduct extensive experiments on two public datasets, to validate the effectiveness of the image features derived from IQA (Section 3.2) and the relevance of fusing several features.First, we present statistics about the datasets we used. Then, we describe the experimental settings: a brief review of state-of-the-art features for news verification and a selection of the best of these textual features as baselines. Finally, we present experimental results and analyze the features to achieve insights into MONITOR.

4.1 Datasets

To evaluate MONITOR's performance, we conduct experiments on two well-established public benchmark datasets for rumor detection, i.e., the MediaEval Verifying Multimedia Use [6] and the FakeNewsNet [26] benchmarks. Their statistics are shown in Table 4.

MediaEval is collected from Twitter and includes all three characteristics: text, social context and images . It is designed for message-level verification. The dataset has two parts: a development set containing about 9,000 rumor and 6,000 non-rumor tweets from 17 rumor-related events; a test set containing about 2,000 tweets from another batch of 35 rumor-related events. We remove tweets without any text or image, thus obtaining a final dataset including 411 distinct images associated with 6,225 real and 7,558 fake tweets, respectively.

FakeNewsNet is one of the most comprehensive fake news detection benchmark. Fake and real news articles are collected from the fact-checking websites PolitiFact and GossipCop. Since we are particularly interested in images in this work, we extract and exploit the image information of all tweets. To keep the dataset balanced, we randomly choose 2,566 real and 2,587 fake news events. After removing tweets without images, we obtain 56,369 tweets and 59,838 images.

Table 4: MediaEval and FakeNewsNet statistics

Dataset	Set	Tweets		Images
		Real	Fake	
MediaEval	Training Set	5,008	6,841	361
	Testing Set	1,217	717	50
FakeNewsNet	Training Set	25,673	19,422	47,870
	Testing Set	6,466	4,808	11,968

4.2 Experimental Settings

Baseline Features We compare the effectiveness of our feature set with the best textual features from the literature. First, we adopt the 15 best features extracted by Castillo *et al.* from aspects of message content, user, topic and propagation tree, to analyze the information credibility of news propagated through Twitter [8]. We also collect a total of 40 additional textual features proposed in the literature [10, 11, 16, 32], which are extracted from text content, user information and propagation properties (Table 5).

Feature Sets The features labeled *Textual* are the best features selected among message content and social context features (Tables 1 and 2). We select them with the information gain ratio method [13], which is commonly used for measuring the goodness of attributes in decision tree learning, over both datasets. It helps select a subset of 15 relevant textual features with an information gain larger than zero (Table 6).

The features labeled *Image* are all the image features listed in Table 3. The features labeled *MONITOR* are the feature set that we propose, consisting of the fusion of textual and image feature sets. The features labeled *Castillo* are the above-mentioned best 15 textual features. Eventually, the features labeled *Wu* are the 40 textual features identified in literature.

Classification Model To assess the robustness of our proposal, we execute various learning algorithms for each feature set. The best results are achieved by four supervised classification models:

Table 5: 40 features from the literature

Feature
Fraction of Question, Exclamation Mark,
Count of Message,
Average Word, Character Length,
Fraction of First, Second, Third Pronouns,
Fraction of URL,@,#,
Count of Distinct URL,@,#,
Fraction of Popular URL,@, #,
Whether the Tweet includes pictures,
Average Sentiment Score,
Fraction of Positive, Negative Tweets,
Count of Distinct People, Location, Organization,
Fraction of People, Location, Organization,
Fraction of Popular People, Location, Organization.
Count of Distinct Users, Fraction of Popular Users,
Count Followers, Followees, Posted Tweets,
Whether the User Has Facebook Link,
Fraction of Verified User, Organization,
Count comments on the original message
Time interval between original message and repost

Table 6: Best textual features selected by gain ratio

MediaEval	FakeNewsNet
Tweet_Length	Tweet_Length
Num_Negsentiwords	Num_Words
Num_Mentions	Num_Questmark
Num_URLs	Num_Uppercasechars
Num_Words	Num_Exclammark
Num_Uppercasechars	Num_Hashtags
Num_Hashtags	Num_Negsentiwords
Num_Exclammark	Num_Possentiwords
Num_Thirdorderpron	Num_Followers
Times_Listed	Num_Friends
Num_Tweets	Num_Favorites
Num_Friends	Times_Listed
Num_Retweets	Num_Likes
Has_Url	Num_Retweets
Num_Followers	Num_Tweets

decision trees, KNNs, SVMs and random forests. We use their Scikit-learn library for Python [23] implementation. Training and validation is performed for each model through a 5-fold cross validation. Note that, for MediaEval, we retain the same data split scheme. For FakeNewsNet, we randomly divide data into training and testing subsets with the ratio 0.8:0.2. Table 7 present the results of our experiments. We use *Accuracy*, *Precision*, *recall* and F_1 score to evaluate the overall prediction performance.

4.3 Classification Results

From the classification results recorded in Tables 7, we can make the following observations.

Performance Comparison With MONITOR, using both image and textual feature allows all classification algorithms to achieve better performance than baselines. Among the four classification models, the random forest generates the best accuracy: 96.2% on MediaEval and 88.9% on FakeNewsNet. They indeed perform 26% and 18% better than Castillo and 24% and 15% than Wu, still on MediaEval and FakeNewsNet, respectively.

Compared to the 15 "best" textual feature set, the random forest improves the accuracy by more than 22% and 10% with image features only. Similarly, the other three algorithms achieve an accuracy gain between 5% and 9% on MediaEval and between 5% and 6% on FakeNewsNet. Compared to the 40 additional textual features, all classification algorithms generate a lower accuracy when using image features only. This is due to the lack of social context and also because textual features are selected from a wide range of textual properties.

Table 7: Classification results

Model	Feature sets	MediaEval				FakeNewsNet			
		Accuracy	**Precision**	**Recall**	F_1	**Accuracy**	**Precision**	**Recall**	F_1
Decision Trees	Textual	0.673	0.672	0.771	0.718	0.699	**0.647**	0.652	0.65
	Image	0.632	0.701	0.639	0.668	0.647	0.595	0.533	0.563
	MONITOR	**0.746**	**0.715**	**0.897**	**0.796**	**0.704**	0.623	**0.716**	**0.667**
	Castillo	0.643	0.711	0.648	0.678	0.683	0.674	0.491	0.569
	Wu	0.65	0.709	0.715	0.711	0.694	0.663	0.593	0.627
KNN	Textual	0.707	0.704	0.777	0.739	0.698	0.67	0.599	0.633
	Image	0.608	0.607	0.734	0.665	0.647	0.595	0.533	0.563
	MONITOR	**0.791**	**0.792**	**0.843**	**0.817**	**0.758**	**0.734**	**0.746**	**0.740**
	Castillo	0.652	0.698	0.665	0.681	0.681	0.651	0.566	0.606
	Wu	0.668	0.71	0.678	0.693	0.694	0.663	0.593	0.627
SVM	Textual	0.74	0.729	0.834	0.779	0.658	**0.657**	0.44	0.528
	Image	0.693	0.69	0.775	0.73	0.595	0.618	0.125	0.208
	MONITOR	**0.794**	**0.767**	**0.881**	**0.82**	**0.704**	0.623	**0.716**	**0.667**
	Castillo	0.702	0.761	0.716	0.737	0.629	**0.687**	0.259	0.377
	Wu	0.725	0.763	0.73	0.746	0.642	0.625	0.394	0.484
Random Forest	Textual	0.747	0.717	0.879	0.789	0.778	0.726	0.768	0.747
	Image	0.652	0.646	0.771	0.703	0.652	0.646	0.771	0.703
	MONITOR	**0.962**	**0.965**	**0.966**	**0.965**	**0.889**	**0.914**	**0.864**	**0.889**
	Castillo	0.702	0.727	0.723	0.725	0.714	0.669	0.67	0.67
	Wu	0.728	0.752	0.748	0.75	0.736	0.699	0.682	0.691

While image features play a crucial role in rumor verification, we must not ignore the effectiveness of textual features. The role of image and textual features is complementary. When the two sets of features are combined, performance is significantly boosted.

Illustration by Example To more clearly show this complementarity, we compare the results reported by MONITOR and single modality approaches (textual and image). The fake rumor messages from Figure 1 are correctly detected as false by MONITOR, while using either only textual or only image modalities yields a true result.

In the tweet from Figure 1(a), the text content solely describes the attached image without giving any signs about the veracity of the tweet. This is how the textual modality identified this tweet as real. It is the attached image that looks quite suspicious.By merging the textual and image contents, MONITOR can identify the veracity of the tweet with a high score, exploiting some clues from the image to get the right classification.

The tweet from Figure 1(b) is an example of a rumor correctly classified by MONITOR, but incorrectly classified when only using the visual modality. The image seems normal and the complex semantic content of the image is very difficult to capture by the image modality. However, the words with strong emotions in the text indicate that it might be a suspicious message. By combining the textual and image modalities, MONITOR can classify the tweet with a high confidence score. This tweet presents a particular type of rumor that is very challenging to identify, because the attached image has been misused from its original context: the two boys were actually photographed in Vietnam in 2007 and have nothing to do with the Nepal Earthquake in 2015.

4.4 Feature Analysis

The advantage of our approach is that we can achieve some elements of interpretability, especially to explain the contribution of image and textual features in the prediction process. Thus, we conduct an analysis to illustrate the importance of each feature set. To identify what features have the most predictive power, we depict the first most 15 important features achieved by the random forest. Variables of high importance are drivers of the classification and their values have a significant impact on the prediction.

Figure 5 shows that, for both datasets, visual characteristics are in the top five features. The remaining features are a mix of text content and social context features. These results validate the effectiveness of the IQA image features issued, as well as the the importance of fusing several modalities in the process of rumor verification.

(a) MediaEval (b) FakeNewsNet

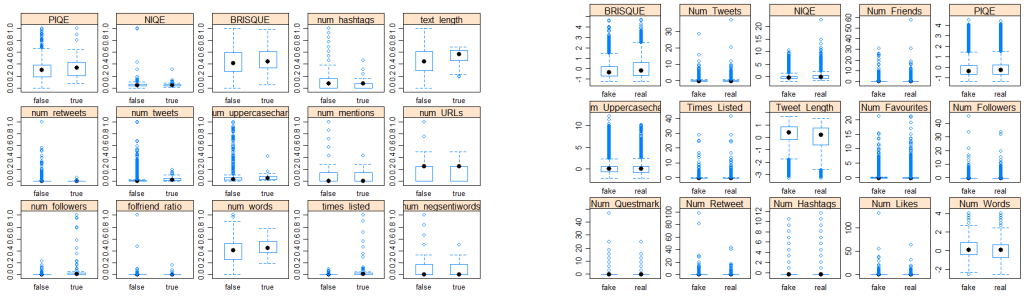
Fig. 5: Top-15 important variables

To illustrate the discriminating capacity of these features, we deploy box plots for each of the 15 top variables on both datasets. Figure 6 shows that several features exhibit a significant difference between the fake and real classes, which explains our good results.

5 Conclusion and Perspectives

To assess the veracity of messages posted on social networks, most machine learning techniques ignore the images attached to messages. In this paper, to improve the performance of the message verification, we propose a multimodal fusion framework called MONITOR that uses features extracted from the textual content of the message, the social context, and also image features have not been considered until now. Extensive experiments conducted on the MediaEval benchmark and FakeNewsNet dataset demonstrated that: 1) the image features that we introduce play a key role in message veracity assessment; and 2) no single homogeneous feature set can generate the best results alone. They also show that with a classification accuracy higher than 96% on MediaEval, and 89% on FakeNewsNet, MONITOR outperforms state-of-the-art machine learning methods.

Our future research includes two directions. First, we currently fuse textual, context, and image features into a single vector, which is called early fusion. By combining classifiers instead, we



(a) MediaEval (b) FakeNewsNet

Fig. 6: Distribution of true and false classes for top-15 important features

also plan to investigate so-called late fusion. Second, deep learning models are capable of learning from representations of both text and images. In particular, recurrent neural networks (RNNs) are widely used in sentence representation and convolutional neural networks (CNNs) are efficient for image representation. Combining RNNs and CNNs could thus be useful for detecting rumors. However, we would like to compare their performance with MONITOR's to study the tradeoff between classification accuracy, computing complexity, and explainability.

References

1. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: A survey. ACM Computing Surveys **51**(2), 32 (2018)
2. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: ICM 2017. pp. 795–816. ACM (2017)
3. Kanske, P., Kotz, S.A.: Leipzig affective norms for German: A reliability study. Brm **42**(4), 987–991, 2010
4. Redondo, J., Fraga, I., Padrón, I., Comesaña, M.: The Spanish adaptation of ANEW (affective norms for English words).Behavior research methods **39**(3), 600–605, 2007
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JmLr **3**(Jan), 993–1022 (2003)
6. Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y., et al.: Verifying multimedia use at mediaeval 2015. In: MediaEval (2015)
7. Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y.: Detection and visualization of misleading content on twitter. IJMIR **7**(1), 71–86 (2018)
8. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: 20th WWW. pp. 675–684. ACM (2011)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Anips. pp. 2672–2680 (2014)
10. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: WWW 2013. pp. 729–736. ACM (2013)
11. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM DM. pp. 153–164. SIAM (2012)
12. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE transactions on multimedia **19**(3), 598–608 (2017)
13. Karegowda, A.G., Manjunath, A., Jayaram, M.: Comparative study of attribute selection using gain ratio and correlation based feature selection. IJ of ITKM **2**(2), 271–277 (2010)

14. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975)
15. Kwon, S., Cha, M., Jung, K.: Rumor detection over varying time windows. PloS one **12**(1), e0168344 (2017)
16. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th DM. pp. 1103–1108. IEEE (2013)
17. Li, J., Li, X., Yang, B., Sun, X.: Segmentation-based image copy-move forgery detection scheme. IEEE Transactions on IFS **10**(3), 507–518 (2014)
18. Maître, H.: From photon to pixel: the digital camera handbook. John Wiley & Sons (2017)
19. Martin, N., Comm, B.: Information verification in the age of digital journalism. In: SLAA Conference. pp. 8–10 (2014)
20. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 ASILOMAR. pp. 723–727. IEEE (2011)
21. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE SPL **20**(3), 209–212 (2012)
22. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing?: understanding microblog credibility perceptions. In: ACM 2012 CSCW. pp. 441–450. ACM (2012)
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. JMLR **12**, 2825–2830 (2011)
24. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th ICCL. pp. 3391–3401. ACL, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1287
25. Ruchansky, N., Seo, S., Liu, Y.: Csi: A hybrid deep model for fake news detection. In: ACM on CIKM. pp. 797–806. ACM (2017)
26. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
27. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: 2018 IEEE MIPR. pp. 430–435. IEEE (2018)
28. Silverman, C.: Verification handbook: An ultimate guideline on digital age sourcing for emergency coverage. EJC (2014)
29. Venkatanath, N., Praneeth, D., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. In: 2015 NCC. pp. 1–6. IEEE (2015)
30. Volkova, S., Jang, J.Y.: Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In: Proceedings WC2018. pp. 575–583. IWWWeb CSC (2018)
31. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: Eann: Event adversarial neural networks for multi-modal fake news detection. In: 24th acm sigkdd. pp. 849–857. ACM (2018)
32. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: 2015 IEEE 31st DE. pp. 651–662. IEEE (2015)
33. Wu, L., Liu, H.: Tracing fake-news footprints: Characterizing social media messages by how they propagate. In: 11th ACM WSDM. pp. 637–645. ACM (2018)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE ICCV (2017)