Rumor Classification through a Multimodal Fusion Framework and Ensemble Learning

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, Bron, Cedex, F69676, France.
[2]Université de Lyon, Lyon 2, Laboratoire Cogitamus, , , , France.

Contributing authors: a.azri@univ-lyon2.fr; cecile.favre@univ-lyon2.fr; nouria.harbi@univ-lyon2.fr; jerome.darmonti@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) 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. Moreover, prior works have used many classical machine learning models to detect rumors. However, several advanced models are not applied although recent studies have proven the potency of the ensemble machine learning approach. To help resolve the identified gaps, we propose in this study: (1) further the state of the art, by first using a set of advanced image features that are inspired from the field of image quality assessment, then, we introduce the Multimodal fusiON framework to assess message veracIty in social neTwORks (MONITOR), which exploits all message features by exploring four individual machine learning models; and (2) demonstrate the effectiveness of the ensemble learning algorithms for rumor detection task by comparing the performance of MONITOR with five developed meta-learning models. Extensive experiments are conducted on two real-world datasets. The

experimental results show that MONITOR can outperform the state-of-the-art machine learning baselines, and all the proposed ensemble models can increase the performance of MONITOR with varying rates.

Keywords: Social networks, Rumor verification, Image features, Machine learning, Ensemble learning

1 Introduction

After more than two decades of existence, social media platforms has attracted a large number of users. 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.

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 generally extract features from two aspects of messages, textual content (Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2018) and the social context (L. Wu & Liu, 2018). 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 play a crucial role in the news diffusion process. For example, in the dataset collected by (Jin, Cao, Zhang, Zhou, & Tian, 2017), 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.

Furthermore, the majority of work of the literature use the extracted features to train a wide range of machine learning (Volkova & Jang, 2018) or deep learning (Wang et al., 2018) methods. However, several unconventional learning models are not applied although recent studies are demonstrating the effectiveness of ensemble learning approach with promising results (Gutierrez-Espinoza, Abri, Namin, Jones, & Sears, 2020).

(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

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. Motivated by the recent research on ensemble learning-based approach to classification problems (Pang, Xue, & Namin, 2016), we develop various meta-learning models to investigate the performance of ensemble learning for rumor classification task.

Our contribution is threefold. 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) (Goodfellow et al., 2014). 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.

Second, we present MONITOR (Azri, Favre, Harbi, Darmont, & Noûs, 2021b), which exploits all types of message features by exploring the performance of four individual machine learning models. This choice is motivated by the fact that these techniques provide explainability and interpretability about the decisions taken.

Third, we demonstrate the benefit of ensemble learning, by developing five meta-learning models (soft and weighted average voting, stacking, blending, and super learner ensemble) as a contribution of the four base-algorithms, and we compare their performance with MONITOR. To the best of our knowledge we are the first to apply mixture of meta-learning models for the rumor detection task.

Eventually, extensive experiments conducted on two real-world datasets shows 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 (Boididou et al., 2015) and the FakeNewsNet dataset (Shu, Mahudeswaran, Wang, Lee, & Liu, 2018), respectively. Furthermore, all meta-learning algorithms can notably increase the performance of MONITOR with different rates.

The rest of this paper is organized as follows. In Section 2, we first review 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. We investigate and discuss the performance of ensemble models in Section 5. Finally, in Section 6, we conclude this study and outline future research.

2 Related Works

Related work can be divided into the following categories: (1) non-image features and image features that are essential for checking the veracity of microblog posts, and (2) background information regarding Ensemble learning models and their usage for rumor classification.

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. To classify a message as fake or real, Castillo *et al.* (Castillo, Mendoza, & Poblete, 2011) capture prominent statistics in tweets, such as count of words, capitalized characters and punctuation. Beyond these features, lexical words expressing specific semantics or sentiments are also counted. Many sentimental lexical features are proposed in (Kwon, Cha, Jung, Chen, & Wang, 2013), who utilize a sentiment tool called the Linguistic Inquiry and Word Count (LIWC) to count words in meaningful categories.

Other works exploit syntactic features, 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.* (K. Wu, Yang, & Zhu, 2015) train a Latent Dirichlet Allocation model (Blei, Ng, & Jordan, 2003) with a defined set of topic features to summarize semantics for detecting rumors.

The social context describes the propagating process of a rumor (Shu, Wang, & Liu, 2018). Social network features are extracted by constructing specific networks, such as diffusion (Kwon et al., 2013) or co-occurrence networks (Ruchansky, Seo, & Liu, 2017).

Recent approaches detect fake news based on temporal-structure features. Kwon *et al.* (Kwon, Cha, & Jung, 2017) 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.* (Morris, Counts, Roseway, Hoff, & Schwarz, 2012) 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 (K. Wu et al., 2015), who define a feature called "has multimedia" to mark whether the tweet has any picture, video or audio attached. Gupta *et al.* (A. Gupta, Lamba, Kumaraguru, & Joshi, 2013) 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.* (Boididou et al., 2015) propose the Verifying Multimedia Use (VMU) task. Textual and image forensics (Li, Li, Yang, & Sun, 2014) 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 (Boididou et al., 2018).

2.3 Ensemble learning algorithms

Ensemble learning refers to the generation and combination of multiple inducers to solve a particular machine learning task. The intuitive explanation for the ensemble methodology stems from human nature. Often, decision making by a group of individuals results in more accurate, useful, or correct outcome than a decision made by any one member of the group. This is generally referred to as the wisdom of the crowd (Surowiecki, 2005). Using ensemble learning, the performance of poorly performing classifiers can be improved by creating, training, and combining the output of multiple classifiers and thus result in a more robust classification. There are three main approaches for developing an ensemble learner (Zhang & Ma, 2012):

- Boosting, often uses homogeneous-base models trained sequentially;
- Bagging(Bootstrap AGGregatING), which often uses homogeneous-base models trained in parallel; and
- Stacking, which uses mostly heterogeneous-base models trained in parallel and combined using a meta-model.

By averaging (or voting) the outputs produced by the pool of classifiers, ensemble methods provide better predictions and avoid overfitting. Another reason that contributes to the better performance of ensemble learning is its ability in escaping from the local minimum. By using multiple models, the search space becomes wider and the chance for finding a better output becomes higher (Sagi & Rokach, 2018).

Recently ensemble learning methods have shown good performance in various applications, including solar irradiance prediction (Lee, Wang, Harrou, & Sun, 2020), slope stability analysis (Pham, Kim, Park, & Choi, 2021), natural

language processing (Sangamnerkar, Srinivasan, Christhuraj, & Sukumaran, 2020), malware detection (D. Gupta & Rani, 2020), traffic incident detection (Xiao, 2019). Compared to other applications, rumor classification using ensemble learning techniques has very few studies in the past.

In (Kaur, Kumar, & Kumaraguru, 2020) authors proposed a multi-level Voting model for the fake news detection task. The study concludes that the proposed model outperforms the other individual machine learning and ensemble learning models. For multiclass fake news detection (Kaliyar, Goswami, & Narang, 2019) used Gradient Boosting ensemble techniques and compare their performance with several individual machine learning models. Finally, (Al-Ash, Putri, Mursanto, & Bustamam, 2019) find that the Bagging approach to detect fake news showed superior performance than SVM, Multinomial Naïve Bayes, and Random Forest.

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

Figure. 2 shows a general overview of MONITOR. It has two main stages: 1) Features extraction and selection. We extract several 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. 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

To better extract 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 (Martin & Comm, 2014) and the verification handbook (Silverman, 2014). 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. 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

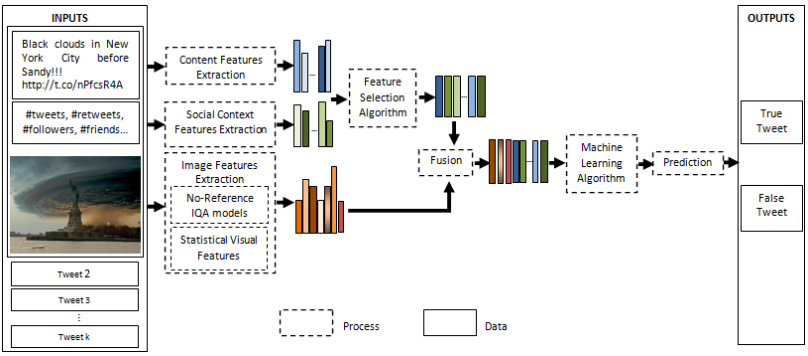


Fig. 2 Overview of MONITOR

performance. Because our focus is the visual features set, we keep all these features in the learning process.

3.2.1 Message Content Features

Content features are extracted from the message's text. We extract characteristics such as the length of a tweet text and the number of its words. It 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. Additional binary features indicate whether the text contains personal pronouns.

We calculate also a readability score for each message using the Flesch Reading Ease method (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975), the higher this score is, the easier the text is to read. 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.

3.2.2 Social Context Features

The social context reflects the relationship between the different users, therefore the social context features 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. 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.

Table 1 Content features

Description
chars, words
(?), (!) mark
uppercase chars
positive, negative words
mentions, hashtags, URLs
happy, sad mood emoticon
1^{st}, 2^{nd}, 3^{rd} order pronoun
The readability score

Table 2 Social context features

Description
followers, friends, posts
Friends/followers ratio, times listed
re-tweets, likes
If the user shares a homepage URL
If The user has profile image
If the user has a verified account
of Tweets the user has liked

3.2.3 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 (Maître, 2017). These degradations are introduced during different processing stages, such as image acquisition, compression, storage, transmission, decompression. 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. 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) (Mittal, Moorthy, & Bovik, 2011), the Naturalness Image Quality Evaluator (NIQE) (Mittal, Soundararajan, & Bovik, 2012), and the Perception based Image Quality Evaluator (PIQE) (Venkatanath, Praneeth, Bh, Channappayya, & Medasani, 2015).

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 compres 22.6603 | Gaussian Noise 28.5840 | Median Blur 41.5620

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

No-reference IQA metrics 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 no-reference IQA for evaluating this type of fake images. Figure 4 displays the BRISQUE score computed for real and fake images generated by image-to-image translation based on GANs (Zhu, Park, Isola, & Efros, 2017).

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 GANs images

Statistical Features.

From attached images, we define four statistical features from two aspects.

Number of Images: 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. 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. We have also a second set of image features composed of

statistical and visual features. 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. 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.

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. 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. Next, we provide the details of both datasets.

4.1.1 MediaEval (Boididou et al., 2015)

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.

4.1.2 FakeNewsNet (Shu, Mahudeswaran, et al., 2018)

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. The detailed statistics of these two datasets are listed in Table 4.

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

4.2.1 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.* to analyze the information credibility of news propagated through Twitter (Castillo et al., 2011). We also collect a total of 40 additional textual features proposed in the literature (A. Gupta et al., 2013; M. Gupta, Zhao, & Han, 2012; Kwon et al., 2013; K. Wu et al., 2015), which are extracted from text content, user information and propagation properties (Table 5).

4.2.2 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 (Karegowda, Manjunath, & Jayaram, 2010). 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.

4.2.3 Build Models

We don't know which model would be good for our problem or what configuration to use. We got an idea from both datasets summarizing that the classes are partially linearly separable in some dimensions. We evaluate a list of mixture of simple linear and non linear algorithms. The best result are achieved by four supervised classification algorithms: Classification and Regression

Table 5 40 features from the literature

Feature
Fraction of (?), (!) Mark,# messages
Average Word, Char Length,
Fraction of 1st, 2nd, 3rd Pronouns,
Fraction of URL,@,#,
Count of Distinct URL,@,#,
Fraction of Popular URL,@, #,
if the Tweet includes pictures,
Average Sentiment Score,
Fraction of Positive, Negative Tweets,
Distinct People, Loc, Org,
Fraction of People, Loc, Org,
Fraction of Popular People, Loc, Org,
Users, Fraction of Popular Users,
Followers, Followees, Posted Tweets,
If the User has Facebook Link,
Fraction of Verified User, Org,
comments on the original message
Time between original message and repost

Table 6 Best textual features selected

MediaEval	FakeNewsNet
Tweet_Length	Tweet_Length
Num_Negwords	Num_Words
Num_Mentions	Num_Questmark
Num_URLs	Num_Upperchars
Num_Words	Num_Exclmark
Num_Upperchars	Num_Hashtags
Num_Hashtags	Num_Negwords
Num_Exclmark	Num_Poswords
Num_Thirdpron	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

Table 7 Configuration space for the hyper-parameters of tested models

Model	Main Hyper-Parameters	Type	Search Space
CART	max_depth	Discrete	[1,21]
	criterion	Categorical	['gini','entropy']
KNN	n_neighbors	Discrete	[1,21]
SVM	C	Discrete	[0.1,2.0]
	γ (RBF kernel)	Discrete	[0.1,1.0]
	Kernel	Categorical	['linear', 'poly', 'rbf','sigmoid']
RF	n_estimators	Discrete	[10,500]
	max_depth	Discrete	[3,20]

Trees(CART), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and random forests(RF). Then, we optimized hyper-parameters of each model (see Table 7) by testing multiple settings using *GridSearchCV* function from Scikit-Learn library for Python (Pedregosa et al., 2011). Subsequently, training and validation is performed for each model through 5-fold cross-validation to obtain stable out-sample results. To implement our models we use Scikit-learn library. 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 8 present the results of our experiments.

4.3 Classification Results

From the classification results recorded in Tables 8, we can make the following observations.

Table 8 Performance of individual machine learning models

Model	Features	MediaEval				FakeNewsNet			
		Acc	**Prec**	**Rec**	F_1	**Acc**	**Prec**	**Rec**	F_1
CART	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**	**0.771**	**0.743**	**0.742**	**0.743**
	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
RF	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

4.3.1 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.

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.

4.3.2 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

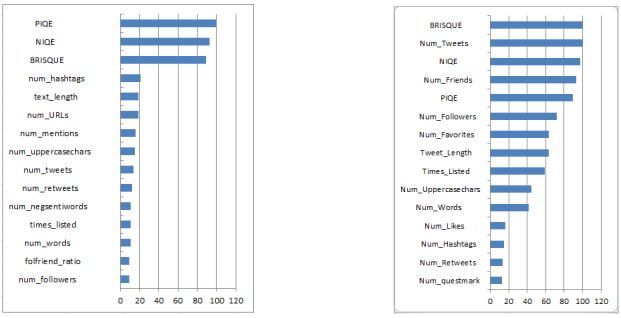
rumor messages from Figure 1 are correctly detected as false by MONITOR, while using either only textual or only image modalities yields a true result.

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.

4.4 Feature Analysis

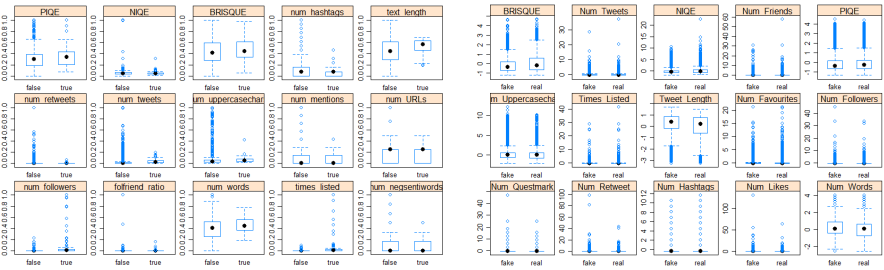
The advantage of our approach is that we can achieve some elements of interpretability. Thus, we conduct an analysis to illustrate the importance of each feature set. We depict the first most 15 important features achieved by the random forest. Figure 5 shows that, for both datasets, visual characteristics 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 Feature importance as indicated by the random forest algorithm

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.

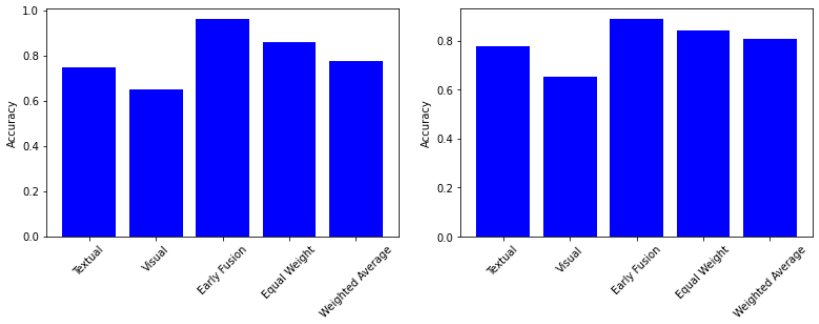


(a) MediaEval (b) FakeNewsNet

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

4.5 Early and Late Fusion

In our previous experiments, we fuse visual and textual features into a single vector in early fusion manner. Another way to merge features is what so-called late fusion. Unlike early fusion, this strategy is based on the combination of classifiers. For this strategy, we train two random forest classifiers by the visual and textual features respectively. To get the final classification results, the predicted probabilities of the two classifiers are combined with equal weights or averaging the weights by feeding the outputs of the two classifiers to a logistic regression model.



(a) (b)

Fig. 7 Performance of Early and Late fusion strategies on: (a) MediaEval (b) FakeNewsNet datasets

From the figure 7, we can see that for both datasets the early fusion method and the two late fusion strategies i.e. equal weight and optimized weight boost the prediction with different rates using separately two sets of features. early fusion has the highest performance score, for both late fusion techniques, equal weight is slightly efficient compared with optimized weight. The performance of late fusion is inferior to that of early fusion because when we train two

models separately on the visual and textual features some dependency between the features will be lost. Practically there is some correlation between the features, for example between BRISQUE and Num_Mention or between PIQE and Text_Length.

5 Ensemble Learning Performance

Applied machine learning often involves fitting and evaluating models on a dataset. Given that we cannot know which model will perform best on the dataset beforehand, this may involve a lot of trial and error until we find a model that performs well or best for our project. This is akin to making a decision using a single expert. Perhaps the best expert we can find. A complementary approach is to prepare multiple different models, then combine their predictions using an ensemble machine learning model.

Because ensemble learning strategies like Bagging and Boosting are typically involves using a single machine learning algorithm (generally a decision tree), we use instead, the stacking strategy or meta learning that seeks a diverse group of members by varying the model types. Figure 8 summarize the key elements of stacking ensemble as follows:

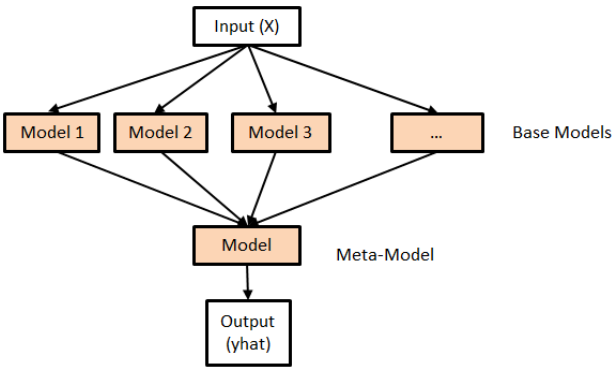


Fig. 8 Stacking Ensemble

- Unchanged training dataset; and
- Different machine learning algorithms (Base Models) for each ensemble member; and
- Machine learning model (Meta-Model) to learn how to best combine predictions.

To measure the performance of ensemble learning models for rumor detection, we develop five meta-models as variants of the stacking strategy.

5.1 Meta-Models

5.1.1 Voting Ensemble

We construct two voting models, (1) a soft voting model that we called $\boldsymbol{MONITOR_{sv}}$ by summing the predictions made by classification models listed in Table 8, and predicting the class label with the largest sum probability, and (2) a weighted average voting model that we called $\boldsymbol{MONITOR_{wav}}$ where model votes are proportional to model performance. The performance of each ensemble model on the training dataset as the relative weighting of the model when making predictions. Performance will be calculated using classification accuracy as a percentage of correct predictions between 0 and 1, with larger values meaning a better model, and in turn, more contribution to the prediction.

5.1.2 Canonical Stacking Ensemble (Wolpert, 1992)

Following the canonical stacking strategy as shown in Figure 8, we construct a model called $\boldsymbol{MONITOR_{st}}$. Concretely, we use three repeats of stratified 10-fold cross-validation of the four classification models to preparing the training dataset (predictions) of the logistic regression meta-model. Furthermore, we train the meta-model on the prepared dataset as well as the original training dataset using 5-fold cross-validation. This can provide an additional context to the meta-model as to how to best combine the predictions.

5.1.3 Blending Ensemble

Is a stacking-type ensemble where the base models are fit on the training dataset and the meta-model is trained on predictions made by each base model on a the validation dataset. At the time of writing this atricle, the scikit-learn library for Python (Pedregosa et al., 2011) does not natively support blending. Instead, we implement a blending model we called $\boldsymbol{MONITOR_{bld}}$ using the scikit-learn models.

5.1.4 Super Learner Ensemble (Van der Laan, Polley, & Hubbard, 2007)

Is a specific configuration of stacking where all base models use the same k-fold splits of the data and a meta-model is fit on the out-of-fold predictions from each model. We can be summarized this procedure in Algorithm 1. Figure 9 below, taken from the original paper, summarizes this data flow. We use MLENS python library (Flennerhag, 2017) to implement the super learner model we called $\boldsymbol{MONITOR_{sl}}$, we splitting the training data into $k = 10$ folds.

Table 9 summarized the results achieved by the best individual machine learning model (RF) and the five stacking algorithms.

Algorithm 1 Super Learner Ensemble Algorithm

1: Select a k-fold split of the training dataset
2: Select m base-models or model configurations
3: **for** each base-model **do**
4: Evaluate using k-fold cross-validation
5: Store all out-of-fold predictions
6: Fit the model on the full training dataset and store
7: **end for**
8: Fit a meta-model on the out-of-fold predictions
9: Evaluate the model on a holdout dataset or use model to make predictions

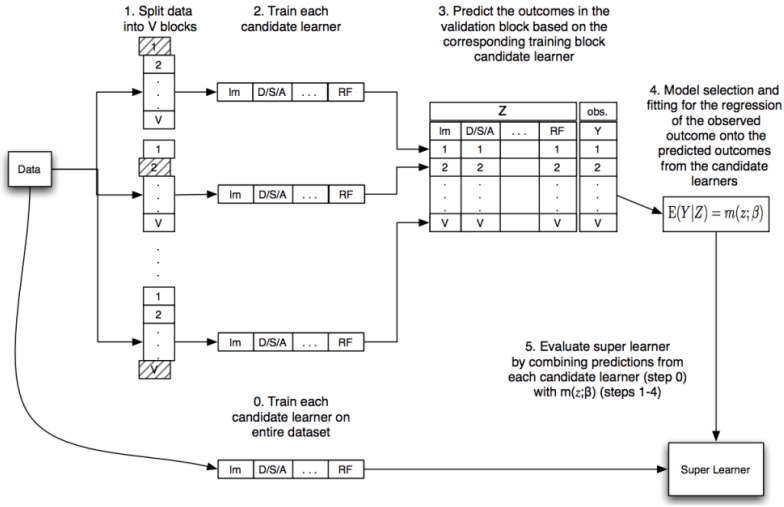


Fig. 9 Diagram Showing the Data Flow of the Super Learner Algorithm (Van der Laan et al., 2007)

5.2 Results Analysis

Comparative analysis of the results shows that all meta-learning models are more efficient than the best individual machine learning model (RF), because by combining multiple models, the errors of a single base-model will likely be compensated by the others, and as a result, the overall prediction performance of the ensemble would be better than that of any single base-model.

For both datasets the canonical stacking algorithm outperforms all models with 98.4% and 93.6% of accuracy on MediaEval and FakeNewsNet dataset respectively. It is because the stacking model takes advantages from the diversity of the predictions made by contributing models. That is, all algorithms are skillful on the classification problem, but in different ways. Figures 10 and 11 illustrates the Box Plot of accuracy scores and the Receiver Operating Curve (ROC) for the canonical stacking ensemble model compared to the standalone

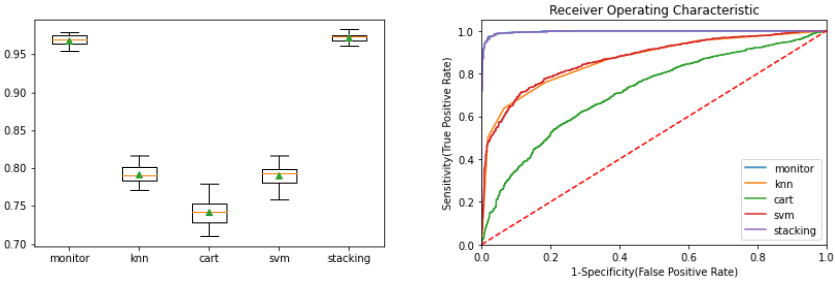
Table 9 Performance of MONITOR and stacking ensemble models

Model	MediaEval				FakeNewsNet			
	Acc	**Prec**	**Rec**	F_1	**Acc**	**Prec**	**Rec**	F_1
MONITOR	0.962	0.965	0.966	0.965	0.889	0.914	0.864	0.889
$MONITOR_{sv}$	0.966	0.955	0.976	0.965	0.897	0.911	0.873	0.892
$MONITOR_{wav}$	0.968	0.968	0.970	0.969	0.906	0.90	0.927	0.914
$MONITOR_{st}$	**0.984**	**0.979**	**0.989**	**0.984**	**0.936**	**0.929**	**0.952**	**0.941**
$MONITOR_{bld}$	0.973	0.975	0.971	0.973	0.915	0.909	0.932	0.921
$MONITOR_{sl}$	0.970	0.980	0.959	0.969	0.921	0.915	0.937	0.926

machine learning algorithms (MONITOR(RF), CART, KNN, and SVM) on both datasets.

Among the five ensemble models, soft voting algorithm achieves the worst results. The reason is a limitation of voting ensemble that it treats all models the same, meaning all models contribute equally to the prediction.

Although the canonical stacking algorithm performed the best, the blending and super learner algorithms achieved scores very close to those of the stacking and therefore turn to be useful for the task of rumor classification.

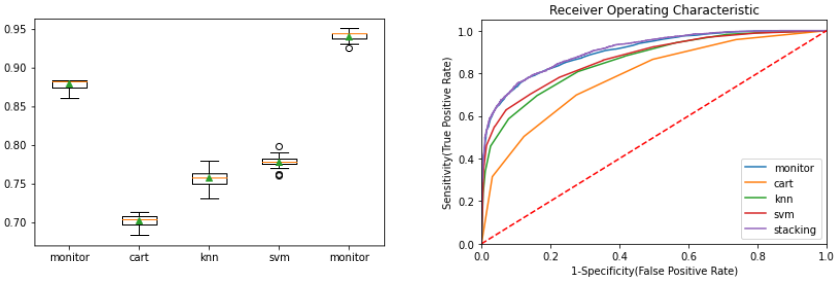


(a) (b)

Fig. 10 Stacking Ensemble model compared to standalone Models for MediaEval dataset: (a) Box Plot of accuracy scores and (b) Receiver operating curve (ROC)

6 Conclusion and Perspectives

To assess the veracity of messages posted on social networks, most of existing techniques ignore the visual content and use traditional machine learning models for classification, although ensemble approach proven promising, and they are considered the state-of-the-art solution for many machine learning challenges. 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. We compare



(a) (b)

Fig. 11 Stacking Ensemble model compared to standalone Models for FakeNewsNet dataset: (a) Box Plot of accuracy scores and (b) Receiver operating curve (ROC)

the performance of MONITOR with five meta-learning ensemble models by combining four base-predictors (KNN, CART, SVM and RF). 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; 2) no single homogeneous feature set can generate the best results alone; and 3) all ensemble algorithms outperforms the best single base-model (RF), canonical stacking acheive best performence on both datasets.

Our future research includes two directions.First, experimenting with other and larger datasets and varying the type, combination, and number of base models for the ensemble. Second, we plan to compare MONITOR performance with a deep learning based approach for rumor classification deepMONITOR (Azri, Favre, Harbi, Darmont, & Noûs, 2021a). Our aim is to study the tradeoff between classification accuracy, computing complexity, and explainability.

References

Al-Ash, H.S., Putri, M.F., Mursanto, P., Bustamam, A. (2019). Ensemble learning approach on indonesian fake news classification. *2019 3rd international conference on informatics and computational sciences (icicos)* (pp. 1–6).

Azri, A., Favre, C., Harbi, N., Darmont, J., Noûs, C. (2021a). Calling to cnn-lstm for rumor detection: A deep multi-channel model for message veracity classification in microblogs. *Joint european conference on machine learning and knowledge discovery in databases* (pp. 497–513).

Azri, A., Favre, C., Harbi, N., Darmont, J., Noûs, C. (2021b). Monitor: A multimodal fusion framework to assess message veracity in social networks. *European conference on advances in databases and information systems* (pp. 73–87).

Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent dirichlet allocation. *JmLr*, *3*(Jan), 993–1022.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., ... others (2015). Verifying multimedia use at mediaeval 2015. *Mediaeval.*

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., Kompatsiaris, Y. (2018). Detection and visualization of misleading content on twitter. *IJMIR*, *7*(1), 71–86.

Castillo, C., Mendoza, M., Poblete, B. (2011). Information credibility on twitter. *20th www* (pp. 675–684).

Flennerhag, S. (2017). mlens documentation.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Anips* (pp. 2672–2680).

Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A. (2013). Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. *Www 2013* (pp. 729–736).

Gupta, D., & Rani, R. (2020). Improving malware detection using big data and ensemble learning. *Computers & Electrical Engineering*, *86*, 106729.

Gupta, M., Zhao, P., Han, J. (2012). Evaluating event credibility on twitter. *Proceedings of the 2012 siam dm* (pp. 153–164).

Gutierrez-Espinoza, L., Abri, F., Namin, A.S., Jones, K.S., Sears, D.R. (2020). Fake reviews detection through ensemble learning. *arXiv preprint arXiv:2006.07912*.

Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, *19*(3), 598–608.

Kaliyar, R.K., Goswami, A., Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. *2019 ieee 9th international conference*

on advanced computing (iacc) (pp. 103–107).

Karegowda, A.G., Manjunath, A., Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *IJ of ITKM*, *2*(2), 271–277.

Kaur, S., Kumar, P., Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), 9049–9069.

Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Kwon, S., Cha, M., Jung, K. (2017). Rumor detection over varying time windows. *PloS one*, *12*(1), e0168344.

Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y. (2013). Prominent features of rumor propagation in online social media. *2013 ieee 13th dm* (pp. 1103–1108).

Lee, J., Wang, W., Harrou, F., Sun, Y. (2020). Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management*, *208*, 112582.

Li, J., Li, X., Yang, B., Sun, X. (2014). Segmentation-based image copy-move forgery detection scheme. *IEEE Transactions on IFS*, *10*(3), 507–518.

Maître, H. (2017). *From photon to pixel: the digital camera handbook*. John Wiley & Sons.

Martin, N., & Comm, B. (2014). Information verification in the age of digital journalism. *Slaa conference* (pp. 8–10).

Mittal, A., Moorthy, A.K., Bovik, A.C. (2011). Blind/referenceless image spatial quality evaluator. *2011 asilomar* (pp. 723–727).

Mittal, A., Soundararajan, R., Bovik, A.C. (2012). Making a "completely blind" image quality analyzer. *IEEE SPL*, *20*(3), 209–212.

Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J. (2012). Tweeting is believing?: understanding microblog credibility perceptions. *Acm 2012 cscw* (pp. 441–450).

Pang, Y., Xue, X., Namin, A.S. (2016). Early identification of vulnerable software components via ensemble learning. *2016 15th ieee international conference on machine learning and applications (icmla)* (pp. 476–481).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *JMLR*, *12*, 2825–2830.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R. (2018, August). Automatic detection of fake news. *Proceedings of the 27th iccl* (pp. 3391–3401). Santa Fe, New Mexico, USA: ACL. Retrieved from https://www.aclweb.org/anthology/C18-1287

Pham, K., Kim, D., Park, S., Choi, H. (2021). Ensemble learning-based classification models for slope stability analysis. *Catena*, *196*, 104886.

Ruchansky, N., Seo, S., Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. *Acm on cikm* (pp. 797–806).

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249.

Sangamnerkar, S., Srinivasan, R., Christhuraj, M., Sukumaran, R. (2020). An ensemble technique to detect fabricated news article using machine learning and natural language processing techniques. *2020 international conference for emerging technology (incet)* (pp. 1–7).

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H. (2018). Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Shu, K., Wang, S., Liu, H. (2018). Understanding user profiles on social media for fake news detection. *2018 ieee mipr* (pp. 430–435).

Silverman, C. (2014). *Verification handbook: An ultimate guideline on digital age sourcing for emergency coverage*. EJC.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Van der Laan, M.J., Polley, E.C., Hubbard, A.E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Venkatanath, N., Praneeth, D., Bh, M.C., Channappayya, S.S., Medasani, S.S. (2015). Blind image quality evaluation using perception based features. *2015 ncc* (pp. 1–6).

Volkova, S., & Jang, J.Y. (2018). Misleading or falsification: Inferring deceptive strategies and types in online news and social media. *Proceedings wc2018* (pp. 575–583).

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. *24th acm sigkdd* (pp. 849–857).

Wolpert, D.H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.

Wu, K., Yang, S., Zhu, K.Q. (2015). False rumors detection on sina weibo by propagation structures. *2015 ieee 31st de* (pp. 651–662).

Wu, L., & Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. *11th acm wsdm* (pp. 637–645).

Xiao, J. (2019). Svm and knn ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and its Applications*, *517*, 29–35.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: methods and applications*. Springer.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 ieee iccv*.