Including Images into Message Veracity Assessment in Social Media

Abderrazek Azri, Cécile Favre, Nouria Harbi, Jérôme Darmont

Including Images into Message Veracity Assessment in Social Media

Abderrazek Azri
a.azri@univ-lyon2.fr

Cécile Favre
cecile.favre@univ-lyon2.fr

Nouria Harbi
nouria.harbi@univ-lyon2.fr

Jérôme Darmont
jerome.darmont@univ-lyon2.fr

University of Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès France,
F69676 Bron Cedex, France

ABSTRACT

The extensive use of social media in the diffusion of information has also laid a fertile ground for the spread of rumors, which could significantly affect the credibility of social media. An ever-increasing number of users post news including, in addition to text, multimedia data such as images and videos. Yet, such multimedia content is easily editable due to the broad availability of simple and effective image and video processing tools. The problem of assessing the veracity of social network posts has attracted a lot of attention from researchers in recent years. However, almost all previous works have focused on analyzing textual contents to determine veracity, while visual contents, and more particularly images, remains ignored or little exploited in the literature. In this position paper, we propose a framework that explores two novel ways to assess the veracity of messages published on social networks by analyzing the credibility of both their textual and visual contents.

CCS Concepts

•**Information systems** → *Social networking sites;* •**Human-centered computing** → *Social network analysis;*

Keywords

Online social networks, Rumors, Veracity, Image forgery detection

1. INTRODUCTION

With the rise of social media platforms, information is generated and propagated at an unprecedented rate. The simplicity of the sharing process has led to a large volume of news content spreading over social networks and reaching vast numbers of users in a short time.

The convenience and openness of microblogs have also fostered the spread of rumors, which have become a serious public concern recently. The spread of rumors is also facilitated by the availability of low-cost and straightforward multimedia content processing tools, such that everyone can easily edit the content of an image or video [16]. Therefore, the automatic assessment of information credibility on social networks becomes a mandatory requirement to limit the propagation of rumors.

On a microblogging platform such as Twitter, users publish short messages (i.e., tweets) that may contain, in addition to "classical" text, tags such as URLs, hashtags or references to other users. Moreover, the message can be accompanied by multimedia content such as an image or video. Consequently, the textual and visual content that forms the message should be considered together. Moreover, the message's veracity depends on the veracity of each of these components.

Most existing approaches for automatic rumor detection are based only on the textual content of messages to predict the veracity of online content. Unlike these approaches, we propose in this position paper a framework that explores the veracity of messages from social networks by analyzing both their textual and visual content in a single process.

For this purpose, we propose two techniques to achieve this challenging task. The first is based on the extraction of appropriate features from the text and useful statistical and visual features from the image. These features are merged to train a supervised classifier to assess the veracity of messages.

The second technique involves a verification system composing two classification tiers. The first classification tier determines the veracity of the textual content based on textual feature extraction; the second classification tier determines the veracity of the image by adopting a forgery detection method. Finally, scores returned from both classifiers are fused to provide the final decision about message veracity.

The rest of this paper is structured as follows. We present the state of the art in Section 2. We detail the definition and categories of fake images circulating on social media in Sec-

INTIS '2019 Tangier, Morocco

DOI: 10.475/123_4

tion 3. We describe the framework we propose in Section 4. We discuss our framework in Section 5. We conclude this paper and present perspectives for our future work in Section 6.

2. RELATED WORKS

We review related work in two main areas: rumor analysis and images forgery detection. Because these are two extensive research areas, we just provide an outline of the research that is most closely related to ours.

2.1 Rumor Analysis in Social Media

Almost existing studies solve the rumor detection problem in feature-based approaches and supervised machine learning scheme. Features from text content [7, 15], users [23], and propagation patterns [32], are extracted to train a classifier on labeled training data. Some recent works further improve the classification result with graph-based optimization methods [10, 12].

To represent high-level abstract semantics in the rumor detection task, the approaches of [14, 32] use features based on topic modeling (LDA). To overcome the limitation of these manually crafted features, *Ma et al.* [19] represents tweets in an event with deep neural networks by using recurrent neural networks (RNN), to learn the representation of tweets in a time series.

Only a few recent studies aim to verify the credibility of multimedia content in addition to text. In [10, 32], some basic features are provided for images attached in tweets. Text and image features are extracted to automatically predict the veracity of a tweet that shares multimedia content in [4].

Finally, *Jin et al.* [11] proposes an RNN with the attention mechanism to fuse features from the text, image, and social context for detecting rumors on microblogs. However, the proposed system in [11] has several limitations. First, this system does not explain the final classification, this problem is one of the weak points in the use of neural networks. Also, the system as it stands, cannot handle the detection of information misuse, this corresponds to a bad association of an image and a text representing two different contexts.

2.2 Image Forgery Detection

Creating forged images by manipulating the original image content is called digital image forgery. This field that intends to verify the authenticity of images has been developed significantly against the problem of image forgeries in many domains like legal services, medical images, and forensics.

There are two main ways to detect alteration: classification and localization. In the classification detection method, the output is binary, whether the image is authentic or forged. On the other hand, the localization method not only detects whether an image is authentic or falsified, but also gives the regions that are manipulated in the image if it's falsified [25, 31].

Image forgery detection methods can be divided into two types: active and passive. An active forgery detection technique uses a known authentication code embedded into the image content such as digital watermarking [18] or digital signatures [28]. Passive forgery detection technique is the process of authenticating images with no requirement of prior information, just the image itself [2, 3].

In passive forgery detection, if a forged image involves parts of more than one image, then resultant image is called a spliced image; if the forged image involves parts of the same image, it is called a copy-move forged image.

3. FAKE IMAGES IN SOCIAL MEDIA

The content of a message on social media is usually linked to an event or a context. For this purpose, the definition of a fake image leads us first of all to define the event.

In literature, the definition of an event can be heterogeneous even though it shares a common characteristic; events are in general said to occur, or happen, meaning that they are entities that unfold over time and/or space [27, 30].

A fake image is defined as any image attached to a message that does not accurately represent the event to which it refers [5].

Following the definition of the event, a fake image is defined as either miscontextualized or tampered. A miscontextualized image presents inconsistencies with at least one aspect of the event, for instance, temporal or geographical misplacement. In other words, it is a bad association of an image and a text representing two different contexts.

Figure 1(a) shows a miscontextulized image that seems to show Paris the day after the attacks of November 2015. However, by searching on Google Image, we can find the origin of the image that dates back to 2008. While the image used to show the hurricane sandy in Figure 1(b) is deliberately tampered .

4. THE PROPOSED FRAMEWORK

The key idea of our proposal is to exploit both textual and visual content of messages to assess the veracity of entire online content by studying the possibilities of text and image processing in a single framework.

Figure 2 presents an overview of the main components of the proposed framework.

4.1 First Technique

The detection of the veracity of messages is formalized in this technique as a classification problem, and it mainly contains two phases: features extraction and model training.

In this method, several useful features are extracted from the message text and the attached image. After features extraction, "traditional" classifiers are built, based on these features to identify messages veracity as real or fake.

4.1.1 Features Extraction

Textual Features Extraction.

Features come from two main aspects of false information or rumor: content and social context in social media, this because the content is associated with a certain social context during spreading of rumor.

Content features are the features extracted from the text. The social context reflects the relationship among different users and describes the propagating process of a rumor, so social context features are the features extracted from the user behavior and the propagation network. These features are suitable to capture the characteristics of rumors.

Depending on their nature, we distinguish three features categories: message content features, user content features, and propagation properties features.

(a)

(b)

Figure 1: Different types of fake images: (a) miscontextulized image; (b) Digitally manipulated image.

1. **Message Content Features.** They take into account information about the content of the messages themselves; this information may or may not be related to social media. The information that is not related to social media includes the presence of exclamation or question marks, the presence of positive or negative sentiment words, and the size of the message. The information that is linked to social media includes the presence of hashtags, URLs, the fact that the message is original or a re-tweet.
2. **User Content Features.** They are the attributes of the user who posted the message, e.g., we assume properties such as the number of friends, followers, the number of tweets the user has authored in the past or registration age.
3. **Propagation Properties Features.** They take into account information about the propagation tree that can be built from tweet and re-tweets of a message, such as the depth of the re-tweet tree.

We propose a set of features to characterize messages in our dataset. Many of these features have been studied in previous works [1, 5, 24, 26].

Patterns of these categories features for rumor messages are obviously different from that of "normal" ones. Table 1 presents a list of features produced for each message.

Image Features Extraction.

To make the difference between real and fake images, the authors of [13] find that, to describe false information, fake images tend to be eye-catching, and visually striking in contrast to real ones. Also, images in the real news are much denser than those in the fake news. Hence, images in fake and real news have visually and statistically distinctive patterns. Therefore, images have an important impact on detecting fake news in microblogs.

For this purpose, we propose a set of features to characterize images, which are divided into two categories.

1. **Statistical Features.** Similar to the statistical features of textual content, some basic statistics of images proved to be distinctive in separating rumors and non-rumors as to mark the occurrence of images in rumor messages, or popular images that gain more retweeting and comments than others.
2. **Visual Features.** Extracted from Image quality assessment (IQA) field. IQA aims to quantitatively represent the human perception of quality. These metrics are commonly used to analyze the performance of algorithms in different fields of computer vision like image compression, image transmission, and image processing [20].

   IQA is mainly divided into two areas of research: (1) Full-Reference evaluation and (2) 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.

   Since in our case, we don't have the original version of the posted image; therefore, the approach that is fitting for our context is no-reference IQA metrics.

   For this purpose, we use three No-reference algorithms demonstrated to be highly efficient: Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [21], Naturalness Image Quality Evaluator (NIQE) [22], and Perception based Image Quality Evaluator (PIQE)[29].

Table 2 presents a list of some features produced for each image.

4.1.2 Model Training

Extracted features are normalized and concatenated, and a classifier model will be trained using the features. A large amount of supervised model can be used such as Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes. At the end of this step, messages are labeled true or false.

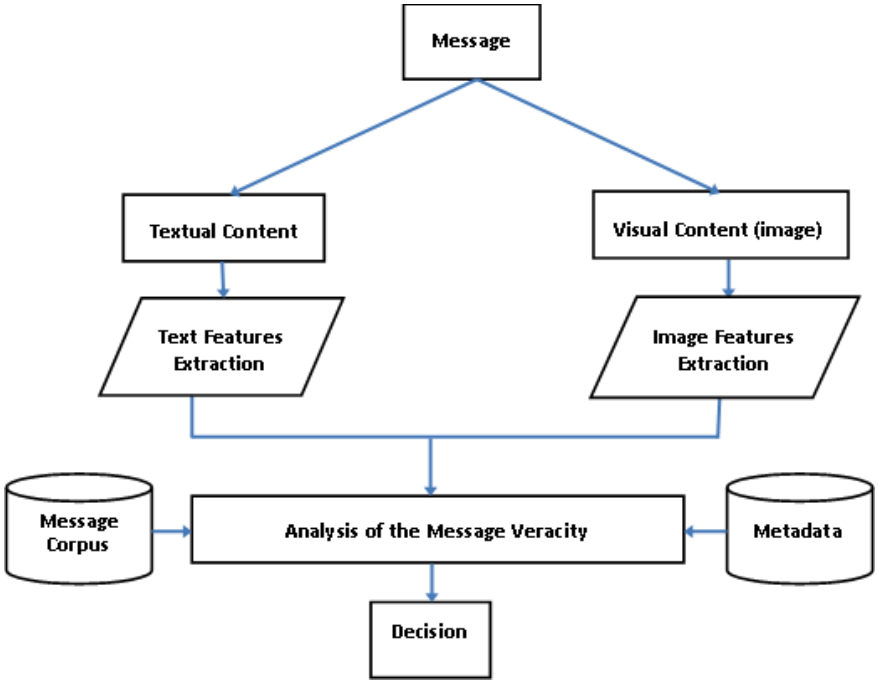

Figure 2: Overview of the proposed framework.

4.2 Second Technique

In this technique, we propose a verification system based on three main steps.

First, we explore the veracity of textual content by adopting the approach followed in the first proposition, using only the features listed in Table 1 and a classifier whose output is the probability of the veracity of textual content of the message, e.g., logistic regression.

Secondly, we explore the veracity of visual content by adopting an image forgery detection method. Almost of the existing forgery detection approaches extract representative and relevant features from an image first, then a suitable classifier is trained and modeled by using the features, and finally, classification performed by using the trained model.

In the field of image forgery detection, no method can detect all forms of falsification, because each technique is designed to identify specific traces of manipulation based on its own assumption; it is, therefore, judicious to fuse multiple outputs from many forgery techniques [9]. Basing on this acknowledgment, we propose to combine numerous classical image forgery techniques on the expectation to detect various image manipulations.

The fusion of these techniques makes it possible to detect more precisely the region that is more likely to contain falsification. Then, relevant statistics are extracted from this region to generate a features vector which will be used to train a classifier, for example, random forests, that provide the probability of image veracity.

Finally, scores returned from the two classifiers are fused to give the final decision.

The final probability of veracity of the whole message is defined as a linear combination of the probability outputs from these two classifiers, as follows:

$$Pr(Message) = \omega_1 \times Pr(Txt_{content}) + \omega_2 \times Pr(Img_{content}) \quad (1)$$

Where, $Pr(Txt_{content})$ and $Pr(Img_{content})$ are respectively the probabilities to be fake for the textual content and the image. The parameters ω_1 and ω_2 are two percentages, where there sum is 100%. We make all possible combinations of ω_1 and ω_2 to evaluate which values are most appropriate to evaluate the authenticity of messages.

The linear combination in equation (1) represents the probability of message to be fake: $Pr(Message)$. If this probability is higher than 50%, then the image is classified as fake, otherwise as real, as can be seen in Figure 3.

5. DISCUSSION

From the analysis of our framework, we can make the following observations and proposals.

The task of distinguishing the credibility of images is challenging, as images can be misleading in many ways. Indeed, in case of miscontextualized images, the vast majority of information misuse requires information from outside the message because it is impossible to detect misuse of information by taking into account only the text and image of a publication.

To overcome this difficulty, and based on the idea that image misused from its context has necessarily got an original context, we propose to search this context on a search engine like Google image[1] or TinEye[2] using this same image. To compare the context of our message with those returned by search engines, we can compare the text of the message with the content of found pages.

[1]https ://images.google.com/
[2]https ://www.tineye.com/

Table 1: List of Features Extracted from Each Message.

Feature category	Description
Message	Length of the tweet Number of Words Contains Question Mark ? Contains Exclamation Mark ! Contains Happy Emoticon Contains Sad Emoticon Contains a personal pronoun in 1st, 2nd, or 3rd person Number of uppercase characters Number of negative sentiment words Number of positive sentiment words Number of mentions The date in which this tweet was written Number of hashtags Number of URLs Retweet count
User	Number of Friends Number of Followers Follower-Friend Ratio Age of user account Is a verified user User has a URL The number of tweets at posting time
Propagation	The depth of a propagation tree The max. size of a level in the propagation tree The degree of the root in a propagation tree

Table 2: List of Features Extracted from Each Image.

Feature category	Description
Visual	BRISQUE score NIQE Score PIQE Score
Statistical	The number of all images in a news event The ratio of the image-tweets in all tweets The ratio of image number to tweet number

In the case where the contexts are not identical, there is then misuse of context either in the request message or in the analyzed page. In this situation, the comparison of the publication dates could make the difference between these cases.

Another possibility is to compare the textual content with the image by extracting the most representative words from the text message [8]. For the image, we can use a system for searching objects and persons in an image [17] which is an active domain, or a bag of visual words technique. Then by comparing the two contents, we can know if, for example, an image represents a person or an object that is not described in the text.

In the field of image forgery detection, the assessment of image veracity depends on the ability of detection methods used to locate alterations. A limitation of these methods is estimating the ability of a modification to make the image fake. Indeed, some modifications will not be intended to mislead the user, such as changing the color of the image. On the other hand, changing a person's face or adding or deleting objects in an image completely changes the user's understanding of the image and therefore its meaning. This aspect related to the threshold for which an image is considered as fake or not has to be further investigated to not send unnecessary warnings to a user who uses an alert system.

6. CONCLUSION AND PERSPECTIVES

Existing approaches for news verification on microblogs ignore image content, which is nonetheless very important, in messages such as tweets. In this position paper, we present and discuss a framework to analyze the veracity of messages in social media.

The framework we propose explores two ways to assess the veracity of messages published on social networks, by analysing both their textual and visual content. The first way is based on the extraction of appropriate features from the text and statistical and visual features from the image. These features are combined to train a supervised classifier to assess the veracity of messages. The second approach presents a verification system with two levels of classification. The first level is used to determine the veracity of textual content; the second determines the veracity of the image by applying a forgery detection method. The returned scores are taken into account to evaluate the final decision on message veracity.

As perspectives for future work, we plan to implement and

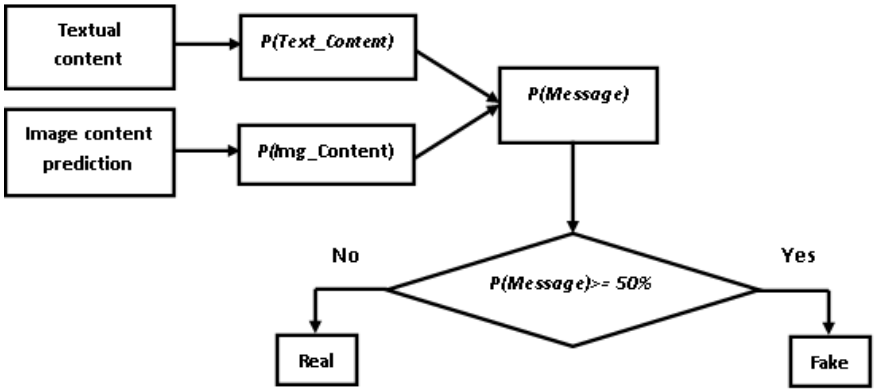

Figure 3: Overview of the second technique.

test our framework with both approaches. However, there are very few standard multimedia rumor detection datasets available. Thus, in addition to the dataset from the MediaEval Verifying Multimedia Use benchmark [6], we intend to build a new dataset for the task. Moreover, to communicate the verification process to end-users, we propose to add a visualization module to preview the results.

7. REFERENCES

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194. ACM, 2008.

[2] K. Asghar, Z. Habib, and M. Hussain. Copy-move and splicing image forgery detection and localization techniques: a review. *Australian Journal of Forensic Sciences*, 49(3):281–307, 2017.

[3] G. K. Birajdar and V. H. Mankar. Digital image forgery detection using passive techniques: A survey. *Digital investigation*, 10(3):226–245, 2013.

[4] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, Y. Kompatsiaris, et al. Verifying multimedia use at mediaeval 2015. In *MediaEval*, 2015.

[5] C. Boididou, S. Papadopoulos, L. Apostolidis, and Y. Kompatsiaris. Learning to detect misleading content on twitter. In *ICMR 2017*, pages 278–286. ACM, 2017.

[6] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 743–748. ACM, 2014.

[7] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW 2011*, pages 675–684. ACM, 2011.

[8] P. Drouin. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115, 2003.

[9] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni. A framework for decision fusion in image forensics based on dempster–shafer theory of evidence. *IEEE Transactions on Information Forensics and Security*, 8(4):593–607, 2013.

[10] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

[11] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *International conference on Multimedia 2017*, pages 795–816. ACM, 2017.

[12] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *ICDM 2014*, pages 230–239. IEEE, 2014.

[13] Z. Jin, J. Cao, J. Luo, and Y. Zhang. Image credibility analysis with effective domain transferred deep networks. *arXiv preprint arXiv:1611.05328*, 2016.

[14] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[15] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[16] F. Lago, Q.-T. Phan, and G. Boato. Visual and textual analysis for image trustworthiness assessment within online news. *Security and Communication Networks*, 2019, 2019.

[17] B. Lavi, M. F. Serj, and I. Ullah. Survey on deep learning techniques for person re-identification task. *arXiv preprint arXiv:1807.05284*, 2018.

[18] C.-Y. Lin and S.-F. Chang. Generating robust digital signature for image/video authentication. In *Multimedia and Security Workshop at ACM Multimedia*, volume 98, pages 49–54. Citeseer, 1998.

[19] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha. Detecting rumors from microblogs

with recurrent neural networks. In *IJCAI*, pages 3818–3824, 2016.

[20] H. Maître. *From photon to pixel: the digital camera handbook*. John Wiley & Sons, 2017.

[21] A. Mittal, A. K. Moorthy, and A. C. Bovik. Blind/referenceless image spatial quality evaluator. In *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*, pages 723–727. IEEE, 2011.

[22] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a âĂIJcompletely blindâĂİ image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[23] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *ACM conference on Computer Supported Cooperative Work*, pages 441–450, 2012.

[24] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[25] M. A. Qureshi and M. Deriche. A bibliography of pixel-based blind image forgery detection techniques. *Signal Processing: Image Communication*, 39:46–74, 2015.

[26] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv preprint arXiv:1011.3768*, 2010.

[27] A. Scherp, T. Franz, C. Saathoff, and S. Staab. F–a model of events based on the foundational ontology dolce+ dns ultralight. In *Proceedings of the fifth international conference on Knowledge capture*, pages 137–144. ACM, 2009.

[28] J.-M. Shieh, D.-C. Lou, and M.-C. Chang. A semi-blind digital watermarking scheme based on singular value decomposition. *Computer Standards & Interfaces*, 28(4):428–440, 2006.

[29] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[30] X.-j. Wang, S. Mamadgi, A. Thekdi, A. Kelliher, and H. Sundaram. Eventory–an event based media repository. In *International Conference on Semantic Computing (ICSC 2007)*, pages 95–104. IEEE, 2007.

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

[32] K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE, 2015.