Deep Learning-Based Rumor Detection on Microblogging Platforms: A Systematic Review

MOHAMMED AL-SAREM[1,2], WADII BOULILA[1,3], (Senior Member, IEEE), MUNA AL-HARBY[1], JUNAID QADIR[4], (Senior Member, IEEE), AND ABDULLAH ALSAEEDI[5]

[1]IS Department, College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia
[2]IS Department, Saba'a Region University, Mareeb, Yemen
[3]RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba 2010, Tunisia
[4]Department of Electrical Engineering, Information Technology University, Lahore, Pakistan
[5]Computer Science Department, College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia

Corresponding author: Wadii Boulila (wadii.boulila@riadi.rnu.tn)

ABSTRACT With the rapid increase in the popularity of social networks, the propagation of rumors is also increasing. Rumors can spread among thousands of users immediately without verification and can cause serious damages. Recently, several research studies have been investigated to control online rumors automatically by mining rich text available on the open network with deep learning techniques. In this paper, we conducted a systematic literature review for rumor detection using deep neural network approaches. A total of 108 studies were retrieved using manual research from five databases (IEEE Explore, Springer Link, Science Direct, ACM Digital Library, and Google Scholar). The considered studies are then examined in our systematic review to answer the seven research questions that we have formulated to deeply understand the overall trends in the use of deep learning methods for rumor detection. Apart from this, our systematic review also presents the challenges and issues that are faced by the researchers in this area and suggests promising future research directions. Our review will be beneficial for researchers in this domain as it will facilitate researchers' comparison with the existing works due to the availability of a complete description of the used performance matrices, dataset characteristics, and the deep learning model used per each work. Our review will also assist researchers in finding the available annotated datasets that can be used as benchmarks for comparing their new proposed approaches with the existing state-of-the-art works.

INDEX TERMS Deep learning, rumor detection, systematic review, Twitter analysis.

I. INTRODUCTION

In recent years, the use of social networking services has rapidly increased. Such social media services constitute an important source of information that can be used in various events. In this paper, we are interested in analyzing information collected from Twitter and Sina Weibo[1] microblogging websites.

In Twitter, users post and interact using tweets (also known as post messages or status), which constitute content units published on Twitter. Initially, when the service was launched, users could only post tweets with a maximum size of 140 characters, but later in 2017, Twitter decided to extend the maximum size to 280 characters, which offers more flexibility to users to express their opinions. Tweets are composed of two main components: the content of the tweet and the user who posted it. Other than tweets, Twitter introduced the ''hashtag'' feature, which categorizes tweets into topics making it easier for users to search and browse tweets for similar content Wyrwoll [64].

Sina Weibo is the most popular and largest micro-blogging website in China which was launched in late 2009. The major differences between Sina Weibo and Twitter with respect to the types of retweeted trending microblogs follow from the fact that for Sina Weibo most of the trends are created due to retweets of media content (such as images, videos, jokes) whereas the trends on Twitter tend to relate more with the current global events and news stories Yang *et al.* [69].

The associate editor coordinating the review of this manuscript and approving it for publication was Jihwan P. Choi.

[1]Often simply referred to as 'Weibo'.

On Twitter, news can be shared without proper restriction or verification leading to a propagation of unconfirmed and unverified statements. Such fake news or rumors can cause public panic and social disturbance. For example, on April 23, 2013, a Twitter account named "Associated Press" falsely claimed that two explosions had occurred in the White House and that President Obama was injured in the attack. A few minutes later White House and "Associated Press" denied the news and announced that the "Associated Press" account has been hacked Chen *et al.* [12]. Similarly, another fake news quoted in Li *et al.* [36] referred to a rumor spread on March 15, 2011, claiming that the prices of salt will rapidly increase after the shortage of salt due to the earthquake and tsunami in Japan. This news triggered panic in people and they started buying salt at high prices.

Vosoughi *et al.* [67] noted that false news is a type of novel information that spreads faster than the truth. This is because people are more likely to share novel information. The extensive spread of rumors or fake news can have a serious negative impact on the individuals and the society such as:

— Rumors can affect the authenticity perception of the news ecosystem leading to a general doubt about all news on Twitter;
— Fake news can change the way people believe or respond to real news;
— Rumors can mislead readers about various events or people. For example, the belief of people on any political party can change the results of the poll Shu *et al.* [54].

A. RUMOR: DEFINITION AND CHARACTERISTICS

In the literature, there are many definitions of rumor. At a very high level, rumor refers to a piece of incorrect information (Cai *et al.* [7]; Liang *et al.* [37]. Wu *et al.* [63] defined a rumor as an unconfirmed statement or false news that often carries malicious information that is created intentionally or unintentionally. Rumors can also be defined as "*unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat, and that function to help people make sense and manage risk*" Difonzo and Bordia [17]. Oxford Dictionary defines a rumor as "*a currently circulating story or report of uncertain or doubtful truth*" ("rumor," n.d.).

Sunstein [56] defines the rumor as "*claims of fact – about people, groups, events and institutions – that have not been shown to be true, but that move from one person to another and hence have credibility not because direct evidence is known to support them, but because other people seem to believe them*". Further discussions on the meaning of a rumor or fake news are given in Cao *et al.* [8], where further sub-categories of rumors such as objective rumor, general rumor, and subjective rumor are introduced. According to the definitions given in Cao *et al.* [8], an 'objective rumor' is strictly equivalent to verified fake information. Once the tweet is confirmed as false by authoritative sources, then

it is labeled as a rumor. 'General rumors' are rumors having unverified truth-values. Similarly, subjective rumors are rumors that their truth-values are determined by the subjective judgments of users.

Some prominent features of rumor propagation in different social networking sites are as follows:

— *Temporal properties* of the rumor spread: according to different social-psychologist theories, rumors can only flourish for a short timeframe and are sometimes spread intentionally to cover the absence of news from the institutional channels Kwon *et al.* [33]. Such information is labeled as *rumor after interrogation*.
— *Structural properties* of the rumors spread: a study on gossips claimed that dense-network structures are less vulnerable to social fragmentation. Kwon *et al.* [33] cited a study on gossips which shows that gossips spread more widely in sparse structures.
— *Linguistic properties* of the rumor spread: a study on laboratory interviews proposed that rumors are expected to be dominated by certain types of sentences such as anxiety, uncertainty, and credulity and outcome relevant involvement Kwon *et al.* [33].

From these points of view, rumors are a particular form of misinformation[2] that are characterized by two features. *First*, the rumors are statements that lack specific standards of evidence. *Secondly*, rumors acquire their power through widespread social transmission Berinsky [6]. Thus, researchers have developed several applications for detecting fake news on Twitter by combining automatic evaluation with the crowd-sourcing annotationGupta *et al.* [23] . After automatic evaluation, a credibility rate is assigned to each tweet by the real-time Credibility Assessment System (TweetCred) as illustrated in Figure 1. Apart from the automatic evaluation, TweetCred also allows users to give their feedback about the rating made by the system. The verified content of the fake news is then labeled based on the aforementioned technique. However, in this method, users are required to give their feedback and human intervention is needed for the detection of rumors. Moreover, the proposed method of rumor detection can only be used to control the spread of false news during propagation instead of detecting the rumor at an early stage.

B. RUMOR DETECTION PARADIGMS

Many rumor detection methods have been proposed in recent years, and most of them are based on machine learning (ML) techniques. The common challenge in most of these works is related to feature extraction from a considered dataset. The task of manual extraction of features requires a lot of time and effort, with a limited efficiency in detecting rumors for most of these works. Recently, deep neural networks have been proposed to simplify the extraction of features and to provide a strong ability for abstract representation learning Li *et al.* [36]. According to Cao *et al.* [8], most of the

---

[2]We use the term "rumor detection" and "false/fake information detection" interchangeably.

FIGURE 1. TweetCred assessment system Gupta *et al.* [23].

automatic rumor detection systems consider the problem of rumor detection as a binary classification task and they may be classified into one of the following three paradigms:

- *Machine learning (ML) (in particular, handcrafted-features-based ML) paradigm:* They apply hand-crafted features to describe the distribution of rumors in high dimensional space. Feature engineering is a crucial requirement for these approaches. These approaches extract features from the textual as well as from visual content based on which classifier is used to separate hyperplane (Castillo *et al.* [9]; Jin *et al.* [27].

- *Networking paradigm*: In contrast to the ML paradigm, the networking paradigm uses several heterogeneous structural social networking features (such as the number of followers, the reply content, timestamp, etc.) along with graph-based optimization methods to evaluate network credibility Jin *et al.* [27]; Jin *et al.* [28]).

- *Deep learning (DL) paradigm*: the approaches of this group are used to learn and fuse multi-modal features automatically. This paradigm, like the ML paradigm, is based on learning from data; but contrary to the handcrafted-features-based approach of the ML paradigm, the DL paradigm does not require any feature engineering since the classifier learns and obtains the required feature during the training phase. The DL paradigm offers many advantages including a significant improvement in performance and the elimination of the cumbersome feature extraction process.

C. CONTRIBUTIONS
Our contributions with this paper are summarized as follows:

- This survey was conducted in a systematic manner guided by the instructions mentioned in Keele [29]. The Systematic Literature Reviews (SLR) methodology aims to provide an assessment of a research subject as fair as possible by being auditable and repeatable. According to Keele [29], the overall purpose of the SLR is to provide a complete list of all studies related to a particular subject while relying on three phases: planning, conducting, and reporting the review. Accordingly, we explain the strategy used for conducting this survey in detail.

- Available datasets used by researchers to validate their approaches in social networks are examined.
- Different architectures of DL for rumor detection are investigated.
- Future research directions in the area of DL for rumor detection are presented.

Since the amount of works that addressed the rumors detection task using the classical machine learning is huge, we limit the current paper to address only works that used DL techniques. Hence, this SLR is the first work that focuses on employing DL techniques in the task of rumor detection.

D. RELATED SURVEYS
The characteristics of rumors have been studied in depth in research surveys. Interested readers can find several surveys that covered the rumor detection in social media from different aspects Zubiaga *et al.* [73]; Cao *et al.* [8]; Kumar and Shah [32]; Zhou and Zafarani [71]. Although the survey of Zubiaga *et al.* [73] presented a deeper overview of the existing detection methods in the context of social media, their work paid less attention to works based on DL algorithms. In Cao *et al.* [8], the authors categorized researchers' efforts into three main paradigms: (i) *machine-learning based approaches*, where the feature extraction is the initial and necessary step to build robust classification algorithms; (ii) *propagation-based approaches,* where the rumors are observed by mining relations among entities; and (iii) *neural network based approaches*.

Kumar and Shah [32] presented a comprehensive survey highlighting how an actor is involved in spreading false information as well as how algorithms are developed to detect false information. Zhou and Zafarani (2018) highlighted some potential characteristics of fake news while also employing DL methods for detection. In summary, most of the published surveys focus generally on rumor detection without focusing on how DL techniques may be employed for rumor detection.

In the current paper, we will present a systematic literature review exploring the most important works related to rumor detection on microblogging such as Twitter and Weibo using only the DL techniques. The proposed survey will also focus on public datasets used in the studied works. Table 1 depicts existing surveys on rumor detection over social media.

E. PAPER ORGANIZATION
The rest of the paper is organized as follows. Section 2 presents a general overview of DL-based methods and architectures. Section 3 details our systematic literature review (SLR) methodology. Section 4 depicts an analysis of the considered studies. Section 5 examines the different architectures of DL used in rumor detection. Section 6 covers the challenges and open issues identified by our survey. Finally, the paper is concluded in Section 7.

II. DEEP LEARNING
The general structure of Artificial Neural Networks (ANNs) is loosely inspired by the way the human brain works.

TABLE 1. Existing surveys on rumor detection for social media.

Research papers	Methods/ Techniques	
	DL	Limitation/ Advantages
(Zubiaga et al. 2017)	P	+Covered different aspects of rumor classification systems; -Mostly covered ML methods and neglects DL methods;
(Cao et al. 2018)	P	-Covered only two categories: RNN-based models and CNN-based models; -Discussed only 2 works per category.
(Kumar and Shah 2018)	X	+Discussed deeply characteristics of rumors and features of false information; -Poorly covered DL methods. Only 2 works are mentioned;
(Zhou and Zafarani 2018)	P	-Briefly discussed DL methods; +More oriented to feature-based methods
(Bondielli & Marcelloni 2019)	P	-Partially covered deep learning methods; -Covered not only detection problem but also verification, stance, and veracity.
(This paper 2019)	√	+The work focuses on detecting rumor using only DL methods; +Follows SLR methodology +Discuses deeply the selected works -Covered only those works that addressed rumor detection and neglects other rumor classification systems

DL: Deep Learning; P: partially covered; X: not covered at all;
+: advantages; -: limitations

ANN represents a computational system consisting of several simple interconnected processes, which aims to solve problems through intensive learning of a large set of training data. The network receives a number of inputs and produces outputs as a network of nodes arranged in layers with connections and weights. The way the nodes are connected and layered defines the architecture of the ANN Grekousis [21]. Usually, an ANN consists of one input layer, one output layer, and one or more layers (called hidden layers) existing between them. The number of hidden layers, connections, and nodes in an ANN architecture are designed according to the complexity of the training data—the more complex the data, the more likely that several additional hidden layers are needed (Grekousis 2019).

Today, deep learning (DL) is very popular in the ML research community due to its superior performance compared to traditional ML in several domains Schmidhuber [53]. DL algorithms belong to the field of ML and more precisely refer to the class of ANNs with many hidden layers. DL methods are characterized by their
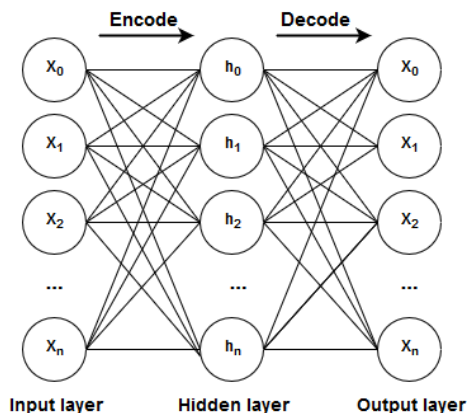

FIGURE 2. AE architecture.

different levels of representation and abstraction to help to make sense of data Deng and Yu [15]. DL algorithms have been introduced into several domains such as natural language processing Collobert and Weston [14], visual data processing Wehrmann et al. [61], speech processing Huang et al. [25], audio processing (Lee et al. [35], and social network analysis Deng et al. [16]. In the literature, several models have been proposed including autoencoder (AE), restricted Boltzmann machine (RBM), deep belief network (DBN), convolutional neural network (CNN), recurrent neural network (RNN), and long short-term memory (LSTM). Interested readers are referred to a more detailed survey of DL algorithms and applications for further reading Pouyanfar et al. [48].

A. AUTOENCODER
Autoencoder (AE) is a DL method that is used to learn features from input data by minimizing the reconstruction error between the input data at the encoding layer and its reconstruction at the decoding layer Vincent et al. [25]. This ensures the learning of important features from data in an unsupervised manner. During the learning process, the encoder maps the data from the input layer to the hidden layer using an encoding function. Then, the decoder maps the encoded values from the hidden layer to the output layer using another decoding function. Figure 2 describes the general architecture of AE. Several types of AE have been proposed—some important types of AE include undercomplete AE, sparse AE, denoising AE, contractive AE, stacked denoising AE, and deep AE.

B. RESTRICTED BOLTZMANN MACHINE
Restricted Boltzmann Machine (RBM) is a generative stochastic neural network, which has two layers of units as shown in Figure 3. The first layer is composed of visible units, whereas the second one is composed of hidden units Salakhutdinov and Hinton [52]. In this type of DL method, the restriction is that there are no connections between the units in a layer; however, a pair of nodes from two different
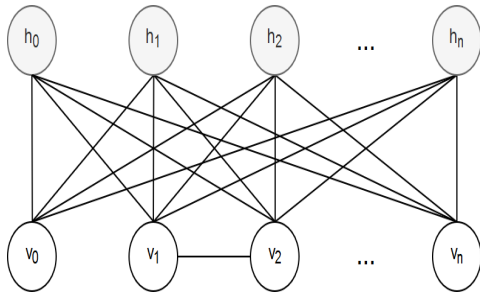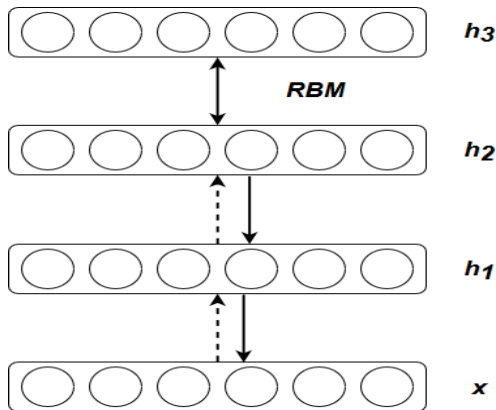
FIGURE 3. RBM architecture.

FIGURE 4. DBN architecture.

FIGURE 5. CNN architecture.

FIGURE 6. RNN architecture.

layers may have a symmetric connection. Units in the first layer represent the observation components whereas units in the hidden layers represent dependencies between observation components. In our context, visible units are text in tweets and hidden units represent dependency between words in the tweets' text. Figure 3 presents the RBM architecture.

C. DEEP BELIEF NETWORK

Deep Belief Network (DBN) is a generative graphical neural network. It is a multi-layer learning architecture with connections between the layers but not between units within each layer. It is composed of several layers of RBM to ensure the pre-training of the model and then uses a feed-forward network during the training step Hinton *et al.* [24]. DBN uses the stack of RBM to extract the hierarchical representation of the training data. The learning process starts by training the RBM using the input data. Then, the output of the first step is used to obtain a representation for the second layer. The second layer is trained using the RBM with the transformed data taken as input. This process is iterated for the desired number of layers to tune all parameters of the DBN architecture with respect to a training criterion. Figure 4 depicts the architecture of the DBN.

D. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) is an unsupervised multilayer feed-forward neural network. It is composed of three stages: namely, convolution, nonlinearity, and
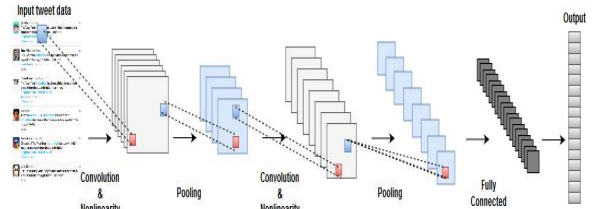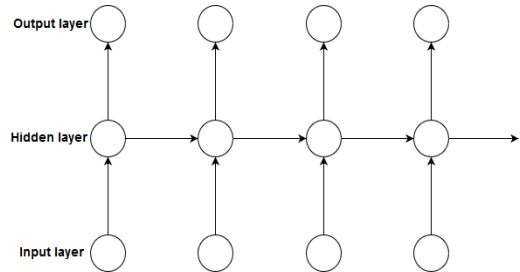
pooling Kim [31]. Figure 5 describes the three stages of the CNN architecture.

— *Convolution*: at this stage, a convolution mathematical-based operation is applied to the input. This operation aims to extract features (called *feature maps*) from the input and then to pass the result to the next layer.

— *Nonlinearity*: the goal of this stage is to include non-linear properties to the network by using a nonlinearity operation such as Rectified Linear Unit (ReLu).

— *Pooling*: the main purpose of this stage is to reduce the dimensionality of the feature maps by applying a function such as max pooling or average pooling.

E. RECURRENT NEURAL NETWORK

Recurrent Neural Network (RNN) is a non-linear adaptive DL network that utilizes a set of sequential information to train the network. In this type of model, connections between nodes form a directed graph along a temporal sequence, which makes it applicable to many domains such as natural language processing, machine translation, video tagging, and speech processing Cho *et al.* 2014. Unlike the feedforward network, the RNN uses recurrent connections to link neurons in the model. It is distinguished by a feedback-loop mechanism to remember past decisions, which is considered as a short-period memory storage. RNN process starts by applying the first hidden layer to input data received by input layers. Then, activations collected by hidden layers are sent to successive hidden layers to produce the output. In the RNN model, each hidden layer is characterized by a weight and a bias. Figure 6 depicts the RNN architecture.

F. LONG SHORT-TERM MEMORY

Long Short-Term Memory (LSTM) uses feedback connections to process an entire sequence of data. It is used in many
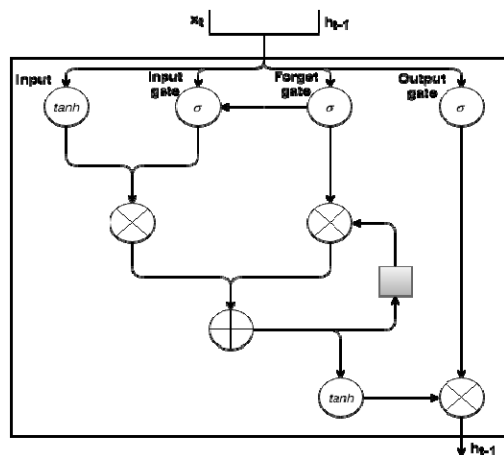
FIGURE 7. LSTM architecture.

fields such as remote sensing, speech and video processing, and handwriting recognition Palangi *et al.* [74] 2016. Applications of LSTM cover classification, processing, recognition, and prediction based on time series data. Figure 7 describes the architecture of LSTM memory cell. The input of a cell in the LSTM network is a variable x_t and the previous cell output h_{t-1}. In addition, LSTM cell has three regulators named input gate, output gate, and a forget gate. The *input gate* decides which values will be updated. The *forget gate* decides which information is to be discarded based on a sigmoid layer. The *output gate* controls the information that is not used at the current time but can be useful in the next steps. These three gates are recurrent and are used to organize information to be saved or discarded.

III. REVIEW METHODOLOGY
The systematic review methodology conducted in this survey follows the instructions mentioned in (Keele 2007). This section is divided into four sub-sections: the first focusing on the development of the review protocol (Section 3.1); the second on the research questions (Section 3.2); the third on the source of information (Section 3.3), and finally on the selection criteria (Section 3.4).

A. DEVELOPMENT OF THE REVIEW PROTOCOL
The first step to conduct this review is to search for relevant research studies in several digital libraries and databases. The number of selected studies is then reduced by applying the inclusion and exclusion criteria. After this, a set of research questions is formulated to thoroughly conduct the proposed study.

B. RESEARCH QUESTIONS
The proposed systematic review aims to deeply analyze how rumor detection systems benefit from applying DL techniques. It also presents a set of tools used to extract data from Twitter and lists the available public datasets used to perform rumor detection.

TABLE 2. Questions considered in our systematic review.

Question	Description
RQ1: How the publications are distributed over the last five years and their types?	The answer to this question gives researchers an overview on the type of publications and their distribution over the countries and years.
RQ2: What is the impact of the studies included in the survey?	Identify citation ratio helps us to include only the relevant works in the area
RQ3: Which datasets are used mostly for conducting rumors analysis?	Identify available datasets helps researchers to use them as benchmarks as well as to compare with their works.
RQ4: Which DL techniques are used most for detecting rumors? Which of these techniques reached the highest performance?	The answer to this question gives researchers an overview of the current trends in applying DL for rumor detection. The second part of this question will aim to help researchers in selecting the baseline for each dataset.
RQ5: Why DL methods are used in the rumor detection field?	With an answer to this question, researchers will be able to understand which DL methods/architectures are suitable for rumor detection.
RQ6: What are the main challenges within the rumor detection field?	The answer to this question helps new researchers to recognize the open research challenges in this area
RQ7: what are the open issues of using DL for rumor detection?	The answer to this question helps new researchers to determine the future path of DL and rumor detection.

Table 2 presents a set of research questions that are used to conduct a systematic literature review.

C. SOURCE OF INFORMATION
To conduct our systematic literature review, the following digital libraries and datasets have been selected:

— IEEE Explore (www.ieeexplore.ieee.org)
— Springer Link (www.springerlink.com)
— Science Direct (www.sciencedirect.com)
— ACM Digital Library (https://dl.acm.org)
— Google Scholar (www.scholar.google.com)

The search keywords used to find relevant studies were "Rumor detection" OR "Fake information" AND "Deep Learning" OR "Deep Neural Networks" OR [name_of_DL_method]. Resulting papers after the search step involve both rumor detection and DL. A preliminary step is needed to remove redundant papers. Details about the number of resulting papers are reported in Table 3.

In addition, 11 works are identified through the references' list and "cited by" searching.

D. SELECTION CRITERIA
Due to a large number of studies that have been founded, we defined a number of restriction criteria to select relevant papers for our survey. These criteria are divided into inclusion and exclusion criteria.

[3]We include papers through Google Scholar only when they are not found by the other identified digital libraries.

Number of returned articles.

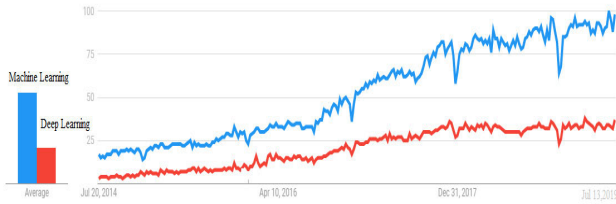| Digital libraries	Number of articles found	Time frame
IEEE Explore	5	
Springer Link	32	
Science Direct	37	2015-2019
ACM Digital Library	33	
Google Scholar	273	
Total	134	

FIGURE 8. Machine learning vs. deep learning for rumor detection (Google trends[4]).

The inclusion criteria is described as follows:

— The literature review covers the period from 2015-2019. This is due to the substantial consideration accorded to DL during the last five years by the research community after the phenomenal progress made by DL in many fields. Figure 8 shows world-wide current research trends in ML domain versus DL domain for social networking analysis. It can be observed that the use of classical ML methods is also common (depicted in red color on Fig. 8) compared to the use of DL, with a notable change in trends starting in 2015.

— The included studies in this systematic review are limited to those works written in English.

— Only studies conducted on Twitter as well as similar microblogs like Sina Weibo are considered. If a study is found in more than one journal or conference proceedings, then the most complete version of the study is included.

In terms of the exclusion criteria, we eliminated studies that are not related to the research questions listed earlier as well as duplicated works or papers published before 2015. Manual screening of the title and abstracts according to the exclusion criteria resulted in the exclusion of 80 studies. In addition, a manual review of the content of remaining studies was performed resulting in 25 related studies. Figure 9 illustrates the procedure for searching for the relevant studies for our SLR.

IV. ANALYSIS OF CONSIDERED STUDIES

This section aims to identify and analyze the resulting relevant papers and provide answers to questions listed in Table 2.
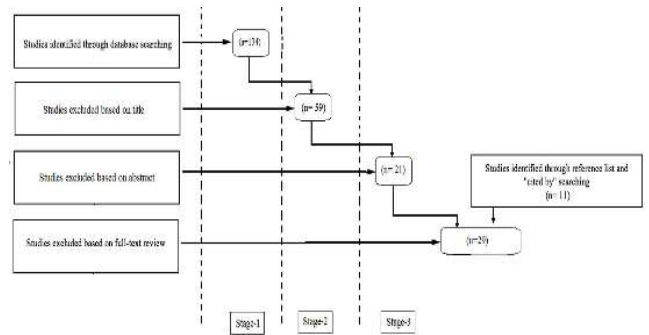
[4]https://trends.google.com. Last access [Jul 13, 2019, 11:20 PM]

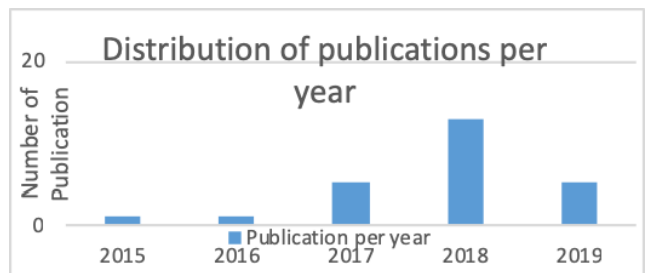**FIGURE 9.** Procedure for searching relevant papers.

FIGURE 10. Publications per year.

To answer the question *RQ1*, the year-wise status and the nature of the sources of publication have been explored as follows:

— Figure 10 shows the number of studies published between 2015 and July 2019. As we can see, there has been a large number of recently published studies due to the increased interest in DL rather than traditional ML for detecting rumors and fake news.

A deeper analysis of the considered studies reveals:

○ A continuously growing body of research related to rumor detection on microblogging platforms over the last 5 years.

○ Figure 11 indicates the number of studies per country. The majority of works were from China with 14 published papers, followed by researchers from America with 4 papers, Australia with 2 papers, while we found only one study from Canada, India, Singapore, Germany, and UK.[5] This implies that the researchers from China, and USA have been paying more attention to the study of rumors proliferation via social media platforms and to the various mitigation strategies that may be adopted.

○ 72% of the considered studies were published in conference proceedings whereas 28% in journals. As we note, the percentage of publications in conferences is greater than the percentage of publications in journals. This can be justified because DL is still not sufficiently familiar in the field of rumor detection. Table 4 presents

[5]Amount of publications per country is computed based on the affiliation of the first author.

TABLE 4. Classification of type of publication for the selected papers. (citation data via google scholar, checked On July 17, 2019.)

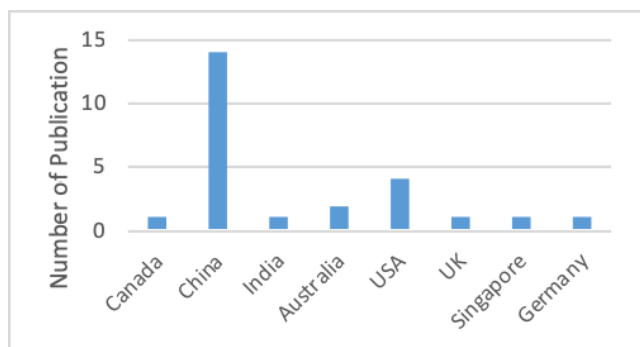| Publication			Type of publication	
Source	Amount of Citations	Key	Journal	Conference
(Alkhudair et al. 2019)	1	K1	√	
(Ma et al. 2019)		K2		√
(Gao et al. 2019)		K3		√
(Geng et al. 2019)		K4		√
(Lui et al. 2019)	2	K5	√	
(Roy et al. 2018)		K6	√	
(Li et al. 2018)		K7		√
(Chen et al. 2018b)	30	K8	√	
(Ma et al. 2018a)	9	K9		√
(Wang et al. 2018)	163	K10		√
(Ajao et al. 2018)	4	K11		√
(Wu et al. 2018)	1	K12	√	
(Xu et al. 2018)	1	K13		√
(Qian et al. 2018)	7	K14		√
(Poddar et al. 2018)		K15		√
(Ma et al. 2018b)	15	K16		√
(Chen et al. 2018a)		K17	√	
(Guo et al. 2018)	5	K18		√
(Jin et al. 2017)	19	K19		√
(Nguyen et al. 2017)	9	K20		√
(Volkova et al. 2017)	56	K21		√
(Yu et al. 2017)	15	K22		
(Ruchansky et al. 2017)	77	K23		√
(Ma et al. 2016)	145	K24		√
(Song et al. 2015)	1	K25	√	

FIGURE 11. Publications per country.

a classification of selected papers according to the type of publication ordered according to the year of publication.

To answer the research question RQ2, we analyze the 25 selected papers in terms of the fields specified in Table 5.

The citation count is an important feature that reflects the quality of the published research study. Table 6 shows that approximately 32% of published research studies have a

TABLE 5. Information analyzed in the considered papers.

Extracted Data	Description
Type of study	Book, journal paper, conference paper
Bibliographic references	Authors, year of publication, title, and source of publication
Citation Number	Citation count that a paper gained over the last 5 years
Dataset	Source, size, and type of the dataset
DL architecture	Type of DL used in the paper
Type of classification	Binary or multi-class classification

TABLE 6. Number of citations for the considered papers over 2015–2019.

Citation Count	≥15	≥ 5 and <15	<5
Number of papers	8	4	13

large audience where the citation count exceeds 15 citations (number of citations is computed based on Google Scholar on July 17, 2019). The highest cited study was the work done by Wang *et al.* [60] which has 163 citations followed by the work of Ma *et al.* [43] with 148 where Ruchansky *et al.* [50] and Volkova *et al.* [66] works have 77 and 56 citations, respectively.

TABLE 7. Available datasets for rumors detection task.

| Dataset | Distribution of dataset | | Format | Provided by | Date | | | Dataset link [9] |
	R	N			C	U	P	
PHEME	1,972 (34.0%)	3,830 (66.0%)	JSON	figshare	Oct 24, 2016	Oct 25, 2016	Oct 24, 2016	https://figshare.com/articles/PHEME_dataset_of_rumors_and_non-rumors/4010619/1
KAGGLE	2,762 (21.2%)	10,233 (78.7%)	CSV	Armineh Nourbakhsh		Mar 27, 2017		https://www.kaggle.com/arminehn/rumor-citation
Newly Emerged rumors in ...	1,01910 (18.4%)	4517 (81.59%)	XLSX	Federal University of Rio de Janeiro			Feb 13, 2019	https://doi.org/10.5281/zenodo.2563864
Ebola 2015			XLSX	Internews	Dec 12, 2015	Jul 17, 2018	May 3, 2017	https://data.humdata.org/dataset/rumors-tracking-liberia-mar-sept-2015
Credibility Corpus			RAR	Nicolas Turenne	Dec 1, 2016		Dec 1, 2016	https://www.data.gouv.fr/es/datasets/credibility-corpus-with-several-datasets-twitter-web-database-in-french-and-english/
BoyerBlaine_threats			XLSX	figshare	Jun 19, 2017		Jun 19, 2017	https://figshare.com/articles/rumor-mongers_as_reliable_sources_Communicating_threat-related_information_suggests_source_reliability_even_for_unlikely_threats/5089867/1
Greece - rumors and Feedback on Migrants -	806 (11.35%)	6293 (88.6%)	XLSX	Internews	Dec 12, 2015	Every six months		https://data.humdata.org/dataset/feedback-and-rumors-migrants-greece-dec-2015-feb-2017

R: rumor; N: Not-rumor; C: Created; U: Updated; P: Published

V. DEEP LEARNING IN RUMOR DETECTION

The study of the considered studies shows that different DL models, frameworks and approaches are proposed for rumor detection. Each study has tested its proposed approach on different datasets to ensure the reliability and credibility of their study. In the following sections, we start by reviewing the most commonly used real-world datasets for the task of rumor detection (a complete list is shown in Table 7). Thereafter, we focus specifically on the datasets used in the considered studies.

A. AVAILABLE PUBLIC DATASET

In this section, we present five datasets namely (1) PHEME; (2) Kaggle; (3) Newly Emerged Rumors; (4) Liberia - Ebola 2015, and (5) Credibility Corpus. (The details of these datasets including download details can be seen in Table 7).

1) PHEME DATASET

The PHEME is a public dataset, which is made by Zubiaga et al. (2016). The dataset is harvested from Twitter by crawling tweets using the Twitter streaming APIs. The dataset

contains both rumors and non-rumors tweets posted during breaking news. According to Zubiaga et al. (2016), breaking news items are categorized into five groups based on the news event as follows:

— Ottawa shooting, which occurred on Ottawa's Parliament Hill in Canada on October 22, 2014.
— Charlie Hebdo shooting, which occurred on the offices of the French satirical weekly newspaper Charlie Hebdo in Paris on January 7, 2015.
— Sydney siege, which occurred in Lindt chocolate cafe, located at Martin Place in Sydney, Australia, on December 15, 2014.
— Ferguson unrest when citizens of Ferguson in Michigan, USA, protested on August 9, 2014.
— Germanwings plane from Barcelona to Düsseldorf crashed in the French Alps on March 24, 2015

The dataset contains, for all five events, 5,802 annotated tweets of which 1,972 were rumors and 3,830 non-rumors. For each event, the authors grouped the related tweets in one directory with two subfolders: a subfolder containing rumors' tweets and the other containing non-rumor s' tweets. Both the folders have subfolders named with a tweet ID. The tweet itself can be found on the 'source-tweet' directory of the tweet in question, and the directory 'reactions' has the set of tweets responding to that source tweet.

2) KAGGLE DATASET
The dataset is available in comma-separated values (CSV) format and consists of three files containing a collection of webpages cited separately in Snopes.com,[6] Emergent.info,[7] and Politifact.com.[8] Researchers can conduct multi-labels classification task easily due to the well-organized structure of the dataset, which is presented as follows:

— Snopes dataset is 16.9k x 12; where the class labels of the rumor assigned by Snopes.com editors are *true*, *false*, *mfalse* (mostly false), *mtrue* (mostly true), *mixture*, or *unverified*.
— Emergent dataset is 2145 x 15; where the class labels of rumors assigned by emergent.inf editors are *true*, *false*, or *unverified*.
— Politifact dataset is 2932 x 12; where the class labels of the rumor assigned by politifact.com editors are *true*, *mostly-true*, *half-true*, *barely_true*, *mostly-false*, *false*, or *pants-fire* ("pants on fire").

3) NEWLY EMERGED RUMORS ON TWITTER
The *newly emerged rumors* dataset is a collection of rumors that rise and fall in a short time. The collection consists of 12 different datasets. The datasets allow researchers to conduct the rumor classification task as well as the rumor tracking task. The class labels of rumor assigned by editors are *true*, *anti-rumor*, a *question about the rumor*, and *post*

not related to the rumor (even though it contains the queries related to the rumor, but does not refer to the rumor).

4) LIBERIA - EBOLA 2015
Ebola-2015 is a collection of rumors occurred during the Ebola crisis in Liberia. The data was collected via mobile phones by a network of Community Health Workers (CHW) and journalists in Liberia between March and September 2015. The data is organized in a way allowing the researchers to conduct a rumor tracking task as well as a rumor detection task.

5) CREDIBILITY CORPUS
The corpus consists of datasets containing information that occurred on social media written in French and English. These datasets are made by crawling the microblogging platform Twitter as well as from the web documents. The corpus consists of: (i) one corpus describing texts from the web database about *rumors and disinformation*; (ii) four corpora from Twitter about *specific rumors* (two each in English and French); (iii) four corpora from Twitter (two each in English and French); and (iv) four corpora from Twitter about *specific rumors* (two each in English and French). The size of the corpus and the amount of collected rumors differ from one corpus to another; however, they share a common column "rumor indicator" with value 1 (is a rumor) or −1 (not a rumor).

B. DATASETS USED IN RESEARCH STUDIES
In this section, we detail the datasets used in the considered papers to track rumors using DL methods.

1) LIAR DATASET
The LIAR dataset is a well-balanced benchmark dataset, which includes 12.8K human-labelled short statements from POLITIFACT.COM's API Wang [59]. The used class label for the truthfulness ratings in the dataset is fine-grained and each statement is categorized into six groups: pants-fire, false, barelytrue, half-true, mostly-true, and true. Figure 12 presents an excerpt from the LIAR datasets listed in Wang [59].

In **K6** and **K12**, an annotated dataset comprising about 12.8K annotated short statements with six fine-grained classes and the information about the authorization source is used. Statements are reported from 2007 to 2016. To enhance the quality of the dataset, Roy *et al.* [57] suggested adding more labelled data as well as the actual statements of the speaker to train the model more accurately. They also recommended using semi-supervised or active learning models for solving such a task. Tracing the actual statements and the information of a speaker's count history of lies leads to get a better understanding of the patterns of the speaker's behavior while making a statement.

---

[6]https://www.snopes.com/
[7]http://www.emergent.info/
[8]https://www.politifact.com/

[9]Last access [13/07/19 at 10:07 AM].
[10]Dataset_R1 (only)

Statement: *"Under the health care law, everybody will have lower rates, better quality care and better access."*
Speaker: Nancy Pelosi
Context: on 'Meet the Press'
Label: False
Justification: Even the study that Pelosi's staff cited as the source of that statement suggested that some people would pay more for health insurance. Analysis at the state level found the same thing. The general understanding of the word "everybody" is every person. The predictions dont back that up. We rule this statement False.

FIGURE 12. An excerpt from LIAR dataset Wang [60].

2) PHEME DATASET

Although the PHEME dataset is a benchmark dataset for stance detection, authors of **K1**, **K2** and **K11** used it to experiment on the rumor detection task. In **K11**, approximately 5800 tweets centered on five rumors stories collected and annotated within the journalism use-case of the project are used. The dataset contains also twitter conversations, which are initiated by fake news. The conversations include tweets and responses of the public on the false news. The dataset contains 330 conversational threads out of which 297 are in English and 33 are German. In **K1**, Alkhodair *et al.* [1] used the PHEME dataset to train the baseline classifiers as well as to check the performance of the proposed LSTM model. They also found that the different feature engineering yields different results per each event of PHEME dataset. In **K2**, the dataset is balanced and the claims with less than 10 tweets are filtered out which did not make in the other works that used PHEME dataset.

3) TWITTER AND WEIBO DATASET

Instead of using available datasets, some researchers prefer to use and collect their own datasets. Social networking like Twitter and Sina Weibo make sample data available to researchers through their APIs Lomborg and Bechmann [39].

Authors of papers **K2**, **K4**, **K8**, **K13**, **K18**, **K23,** and **K24** tested their approaches on available Twitter dataset or Sina Weibo dataset as is without any preparation, whilst others such in **K5**, **K9**, **K10** and **K25** made some filtration or enrich the dataset to conform an acceptable class balancing. The Twitter dataset contains 498 rumors, which are collected using different keywords extracted from the real-time rumor-debunking website namely Snopes.com. This dataset also contains 494 non-rumors events, while, Weibo dataset contains 2,313 rumors and 2,351 non-rumors events—for more details, interested readers are referred to Ma *et al.* [44]. In K10, Wang *et al.* [58]used the dataset collected by Boididou et al. for the Verifying Multimedia Use task that takes place as part of the 2015 MediaEval Benchmark Boididou *et al.* [5]. The data is coming from Twitter as well as from Weibo. The dataset, which is collected from Twitter, is used for detecting fake content. For this purpose, the authors kept textual content as well as the image/video

attached with the tweets. Tweets without any text or image are removed. The Weibo dataset is collected from authoritative news sources of China from May 2012 to January 2016 Jin *et al.* [26]. To enhance the quality of the dataset, the authors followed the preprocessing steps described in (Ma *et al.* [43]; Wu *et al.* [63], which also involved removing duplicated and low-quality images.

In **K3**, **K7** and **K12**, the authors preferred to collect and construct their own corpus. Although this approach is acceptable in the research community, there is no guarantee that the quality of constructed corpus is high and later this will complicate the comparison made by other researchers. In **K7**, Li *et al.* [36] made their dataset by obtaining a set of known rumors from Weibo through Weibo APIs. The dataset has the same number of related microblogs to ensure the balance of the dataset. Accordingly, the dataset contains the same number of rumors and non-rumors events. In **K12**, the dataset comprises about 40K microblogs that are approximately balanced: 9600 true information, 8000 rumors, 8000 biases, 8000 fake news, and 8000 spams.

Another interesting dataset obtained by Weibo APIs is found in **K19**. Jin *et al.* [26] built their corpus by attached to the original tweet text the images and available surrounding social contexts from the rumor and non-rumor sources. At the end, the corpus contains about 40k tweets with images.

4) NEWS ARTICLES' DATASET

To test their proposed framework on both real news articles and corresponding user responses, Qian *et al.* [49] in **K14** conduct experiments on Weibo dataset Ma *et al.* [44] as well as a dataset of news articles, which contains an average length of 950 words.

In **K20**, Nguyen et al. [46] constructed their dataset by tracking stories from online rumors tracking websites such as snopes.com and urbanlegends.about.com using the same process of dataset construction as proposed by Gupta *et al.* [23]. In total, Nguyen *et al.* [46] crawled 4300 stories, including 270 rumors with high impact. For non-rumor events, authors used a corpus made by McMinn *et al.* [45], which covers around 500 real-world events occurred from October 10, 2012, to November 7, 2012. The used corpus is verified manually by the authors. Only the events with the highest number of tweets are kept resulting in a corpus comprising only 230 events. Later another 40 news events happened around the time of rumors are added.

In **K21**, Volkova *et al.* [66]relied on several public resources to check Twitter accounts of suspicious news as well as their corresponding websites. In total, the dataset comprises 174 suspicious news accounts. Table 8 summarizes the datasets used in the considered studies.

5) OTHERS (RUMOR TRACKING WEBSITES)

In **K9**, Ma *et al.* [42] crawled stories posted until March 2015 from two existing rumor-tracking websites, namely snopes.com and emergent.info. As a result, the authors collected 2,299 stories, then, stories that were either

TABLE 8. Datasets used in the selected research studies.

Pub.	Source	Dataset details	Dataset Type		Content Type		Dataset Link
			B	Imb.	T	I	
K1	PHEME dataset & Hand-crafted	PHEME Dataset (section 5.2.2); Thirty-seven tweets about breaking news under hashtag *#WhereAreTheChildren*.		✓	✓		https://figshare.com/articles/PHEME_dataset_of_rumors_and_non-rumors/4010619.
K2	Twitter & PHEME	-Twitter – 498 rumors and 494 non-rumors events -Weibo – 2,313 rumors and 2,351 non-rumors events	✓		✓		
K3	Weibo & Hand-crafted	- Chinese Rumor Corpus (CRC) contains 2,542 rumors; Weibo dataset (2,313 rumors and 2,351 non-rumors events)	✓		✓		
K4	Weibo dataset presented in Ma et al. (2016)	2,313 rumors and 2,351 non-rumors events	✓		✓		
K5	Weibo dataset presented in Ma et al. (2016)	1623 rumors and 1756 non-rumors		✓	✓		
K6	Politifact.com	LIAR		✓	✓		https://www.cs.ucsb.edu/~william/data/liar_dataset.zip

T: Textual; I: Image; B: Balanced; Imb.: Imbalanced;

TABLE 8. *(Continued.)* Datasets used in the selected research studies.

Pub.	Source	Dataset details	Dataset Type Content Type				Dataset Link
			B	Imb.	T	I	
K7	Weibo	250 rumor and 250 non-rumor events	✓		✓		https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0&file_subpath=%2Frumor_detection_acl2017
K8	Twitter & Weibo	-) Twitter – 498 rumors and 494 non-rumors events -) Weibo – 2,313 rumors and 2,351 non-rumors events	✓		✓		
K9	Twitter	94 true stories and 446 false stories		✓	✓	✓	
K10	Twitter & Weibo	Twitter – 7,898 rumors, 6,026 non-rumors events and 514 images Weibo – 4,749 rumors, 4,779 non-rumors events and 9 528 images	✓	✓	✓	✓	https://github.com/MKLab-ITI/image-verification-corpus/
K11	Twitter	5,800 tweets centered on five rumor stories		✓	✓	✓	https://figshare.com/articles/PHEME_dataset_of_rumors_and_non-rumors/4010619/1
K12	Weibo	LIAR Dataset and Weibo – 9,600 True information, 8,000 rumors, 8,000 Fake news and 8,000 Spams	✓	✓	✓		

T: Textual; I: Image; B: Balanced; Imb.: Imbalanced;

TABLE 8. *(Continued.)* Datasets used in the selected research studies.

Pub.	Source	Dataset details	Dataset Type B	Dataset Type Imb.	Content Type T	Content Type I	Dataset Link
K13	Weibo	2,313 rumor and non-rumors 2351 events		✓	✓		http://alt.qcri.org/semeval2017/task8/index.php?id=data-and-tools
K14	Weibo	2,313 rumor and non-rumors 2351 events		✓	✓		Not available yet
K15	Twitter	SemEval-2017 Task 8 dataset	✓		✓		
K16	Twitter	Two datasets, which respectively contain 1,381 and 1,181 propagation trees.		✓	✓		https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0&file_subpath=%2Frumor_detection_acl2017
K17	Twitter	498 rumors and 494 non-rumors	✓		✓		
K18	Twitter & Weibo	Twitter – 498 rumors and 494 non-rumors events; Weibo – 2,313 rumors and 2,351 non-rumors events	✓		✓	✓	https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0&file_subpath=%2Frumor_detection_acl2017
K19	Hand-crafted	-Weibo dataset with 40k tweets only (4749 of them rumors and 4779 non-rumors); -Twitter (9000 rumors and 6000 non-rumors)		✓	✓	✓	
K20	online rumor tracking websites	270 rumors and 270 non-rumors events	✓		✓		

T: **Textual**; I: **Image**; B: **Balanced**; Imb.: **Imbalanced**;

TABLE 8. *(Continued.)* **Datasets used in the selected research studies.**

Pub.	Source	Dataset details	Dataset Type		Content Type		Dataset Link
			B	Imb.	T	I	
K21	Twitter	174 suspicious news accounts and 252 verified news accounts		✓	✓		
K22	Twitter & Weibo	Not reported					
K23	Twitter and Weibo dataset	Twitter – 498 rumors and 494 non-rumors events Weibo – 2,313 rumors and 2,351 non-rumors eve	✓		✓		
K24	Twitter & Weibo	Twitter – 498 rumors and 494 non-rumors events Weibo – 2,313 rumors and 2,351 non-rumors events	✓		✓		
K25	Twitter & Weibo	Three different dataset with different size		✓	✓		

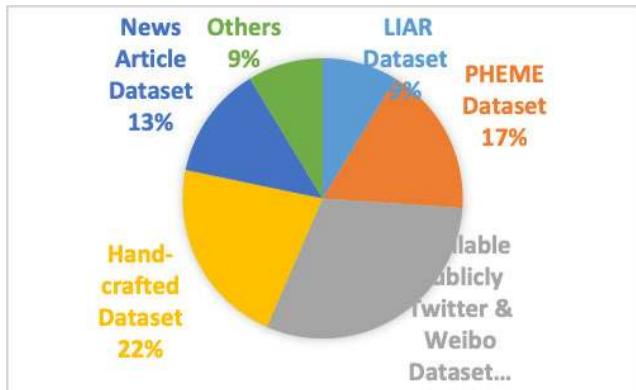T: **Textual;** I: **Image;** B: **Balanced;** Imb.: **Imbalanced;**

FIGURE 13. Distribution of dataset types used in considered studies.

not 'newsworthy' or that did not have an explicit confirmation of veracity are eliminated. The final dataset consists of 94 true and 446 false stories.

In **K17**, Chen *et al.* [12] used two public datasets Kwon *et al.* [33]; Castillo *et al.* [9], and their own dataset containing 498 rumors collected by using keywords extracted from verified fake news published on Snopes.com.

6) CONCLUSION

To answer the question RQ3, the section above provides detailed information about the existing datasets that are intensively used for conducting rumors detection task. Figure 13 shows that, besides, the available publicly datasets (56%), some researchers prefer creating their own datasets by crawling Twitter or Weibo microblogs (22%). 26% of researchers use public PHEME dataset and LIAR datasets; however, these datasets are, generally, used as benchmarks for comparing their proposed approaches. We refer this to the high quality of annotation inter-agreement ratio of PHEME and LAIR datasets. In 13% of cases, researches propose to use news articles' datasets and in 9% use rumor tracking websites.

C. DEEP LEARNING TECHNIQUES

This section is devoted to answering the research questions RQ4 and RQ5. Thus, we start by highlighting the purpose of using DL in rumor detection. Then, we detail the key difference between DL and conventional machine learning techniques. After that, we present the distribution of DL techniques for the rumor detection task. Finally, we provide a general discussion and the main findings obtained from the considered studies.

1) PURPOSE OF USING DL

Compared to conventional ML classifiers (e.g., Naïve Bayes, SVM, Decision Trees), DL performs computations with multiple processing layers for learning data representations with multiple levels of abstraction without feature engineering (Lecun *et al.* [34] In contrast, conventional ML classifiers are dependent on feature engineering, which is usually time-consuming and labor-intensive. The research community,
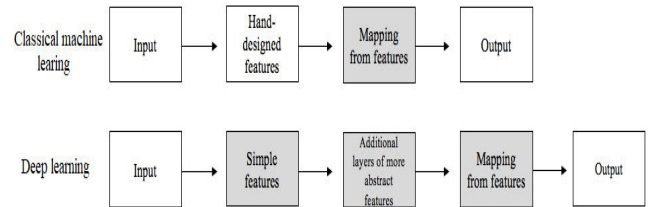

FIGURE 14. Classical machine learning vs. deep learning.
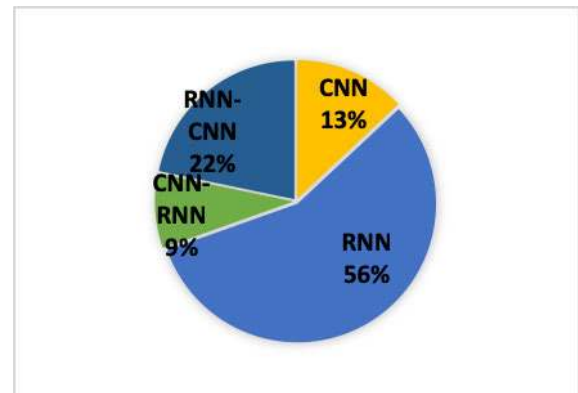

FIGURE 15. Distribution of DL architectures for the considered studies.

therefore, has become more interested in using DL for detecting rumors to avoid the encumbrance of using handcrafted features. In addition, DL can capture more hidden meaningful features comparing to classical machine learning methods (Nguyen et al. 2017). A flowchart illustrating how classic ML and DL relate to each other within different AI disciplines is shown in Figure 14 (Source: Bengio *et al.* [4], with the shaded boxes indicating components that can learn from data.

2) DEEP LEARNING ON RUMOR DETECTION FIELD

To answer the question RQ4, we start by reviewing the DL methods applied in the research studies and their architecture, libraries and tools, and performance metrics. Then, we present general discussion and the main findings obtained from the selected studies. Table 9 summarizes the DL architecture of the proposed rumors detection systems found in the considered research studies. It also identifies the used frameworks, tools, and libraries, and the performance obtained after applying the DL method.

Figure 15 helps to answer the research question RQ4. It shows that researchers mostly prefer to use an RNN architecture (56%) for rumor detection. The CNN architecture also gains good attention with (13%). Although pure CNN or pure RNN shows good performances in fields such as text classification Zhou *et al.* [72], researchers tend to combine it with other architectures for rumor detection to empower the training process. The combination of RNN-CNN comes next in popularity (22%) and then CNN-RNN (9%). In contrast, we noted that AE, DBN and RBM architecture are missing in any research studies.

TABLE 9. Summary about dl architectures, tools/libraries and performance matrices used in research studies.

Pub.	Deep Learning Method		Activation Function	Best Values (Performance metrics)				Tool/ Library	Description of Architecture
	CNN	RNN		Acc.	Pre.	Re.	F		
K1		✓	tanh				0.795	JetBrains IntelliJ IDEA; Deeplearning4j; Word2vec	The standard vanilla RNN model is combined with LSTM to learn long-distance temporal dependencies
K2	✓	✓	softmax	T:0.863; P:0.781	T:0.885; P:0.791	T: 0.892; P: 0.796	T: 0.866; P:0.784		Generative Adversarial Networks (GAN) with GRU and CNN
K3		✓	tanh; softmax		CRC: 0.913; W:0.872	CRC: 0.941; W: 0.921	CRC: 0.926; W:0.896	Word2vec	Three models are proposed, namely RNN, LSTM, and Bi-GRU
K4		✓	tanh; softmax	0.965	0.983	0.982	0.966	Jieba; PyTorch;	RNN with bidirectional GRU and a self-attention layer
K5	✓	✓	-	0.948	0.946	0.958	0.946	LibSVM, JAVA, SkipGram	LSTM network coupled with pooling operation of CNN
K6	✓	✓	ReL	-	0.55	0.45	0.43	Keras, NLTK, Numpy Pandas Sklearn	Models based on CNN and Bi-LSTM networks. The representations obtained from these networks are fed into an MLP for the final classification
K7		✓	tanh	0.889	-	-			Deep Bidirectional Gated Recurrent Unit (D-Bi-GRU)

P: PHEME dataset **Acc.:** Accuracy; **Pre:** Precision; **R:** Recall; **F:** F-score; **CRC:** Chinese Rumor Corpus; **W:** Weibo; - : not used/ does not clear

TABLE 9. *(Continued.)* Summary about dl architectures, tools/libraries and performance matrices used in research studies.

Pub.	Deep Learning Method		Activation Function	Best Values (Performance metrics)				Tool/Library	Description of Architecture
	CNN	RNN		Acc.	Pre.	Re.	F		
K8		✓	tanh	-	T:0.8863 W:0.871	T:0.8571 W:0.8634	T:0.8715 W:0.8672	Theano	LSTM model with soft attention mechanism
K9		✓	Softmax	-	-	-	0.847	Weka, LibSVM, Theano	Gated Recurrent Unit used for representing hidden units
K10	✓		ReLu	T: 0.715 W: 0.827	T: 0.822 W: .847	T: 0.638 W: 0.843	T: 0.719 W: 0.829	Keras	CNN is used to automatically extract features from both textual and visual content of posts
K11	✓	✓	-	0.82	0.443	0.406	0.4059	-	Hybrid models of convolutional neural networks and long-short term recurrent neural network
K12	✓	✓	tanh	T: 0.337 W: 0.433	-	T: 0.597 W: 0.749	T: 0.431 W: 0.549	word2vec; Jieba	RNN-CNN
K13		✓	-	0.944	0.949	0.948	0.944	-	Merged Neural Networks based on LSTM

Acc.: Accuracy; **Pre:** Precision; **R:** Recall; **F:** F-score; **T:** Twitter; **W:** Weibo; - : not reported/ does not clear

TABLE 9. *(Continued.)* Summary about dl architectures, tools/libraries and performance matrices used in research studies.

Pub.	Deep Learning Method		Activation Function	Best Values (Performance metrics)				Tool/ Library	Description of Architecture
	CNN	RNN		Acc.	Pre.	Re.	F		
K14	✓		-	T:0.8984; W:0.88.84				Pre-trained word embedding	Tow level CNN captures semantic information from text, and URG generates responses to new articles to assist in fake news detection.
K15	✓	✓	Softmax, tanh	0.7986	-	-	-	-	CNN and RNN with attention mechanism
K16		✓	Softmax	-	-	-	0.821	Weka, LibSVM, Theano	Tree-structured RNN
K17		✓	logistic sigmoid	-	$\cong$0.68	$\cong$0.69[11]	0.68	tf-idf	Three layer of LSTM architecture for better long-range dependency learning
K18		✓	tanh	T: 0.84.4 W: 0.943	T: 0.948 W: 0.946	T: 0.78 W: 0.94	T: 0.825 W: 0.943	SkipGram Adam-Optimizer	hierarchical neural network combined with social information (HSA-BLSTM)
K19		✓	ReLu	T: 0.682 W: 0.788	T: 0.780 W: 0.862	T: 0.615 W: 0.686	T: 0.689 W: 0.764	Word2Vec VGGNet	RNN with attention mechanism
K20	✓		softmax	0.8119	-	-	-	Scikit-learn TensorFlow Keras	CNN

Acc.: Accuracy; **Pre:** Precision; **R:** Recall; **F:** F-score; **T:** Twitter; **W:** Weibo; **-** : not used/ does not clear

TABLE 9. *(Continued.)* Summary about dl architectures, tools/libraries and performance matrices used in research studies.

Pub.	Deep Learning Method		Activation Function	Performance Metrics				Tool/Library	Description of Architecture
	CNN	RNN		Acc.	Pre.	Re.	F		
K21	√	√	sigmoid softmax	0.95	0.99	-	0.92	Keras, GloVe ADAM, Doc2Ve, TFIDF	linguistically-infused neural network models based on LSTM-CNN model
K22	√		tanh	T:0.777 W:0.933	T:0.820 W:0.92	T:0.848 W:0.945	T:0.793 W:0.932		CNN used paragraph vector to automatically obtain key features of both misinformation and truth information.
K23		√	tanh	T:0.892 W:0.953	-	-	T:0.894 W:0.954	Adam optimaizer	The model consists of two n parts: a module for extracting temporal representation of news articles, and a module for representing and scoring the behavior of users.
K24		√	tanh	T: 0.881 W: 0.91	T: 0.857 W: 0.956	T: 0.968 W: 0.963	T: 0.898 W: 0.914	Weka LibSVM Theano	Three RNN models namely, basic tanh-RNN, 1-layer LSTM/GRU + embedding, and 2-layer GPU+ embedding
K25	√		ReLU	T:0.721 W:0.947	T:0.642 W:0.94	T:1.00 W:0.934	T:0.762 W:0.944		CNN architecture

Acc.: Accuracy; **Pre:** Precision; **R:** Recall; **F:** F-score; **T:** Twitter; **W:** Weibo; **-** : not used/ does not clear

As we can note, although applying LSTM model yields a satisfactory result in most studies, the results are significantly improved when it is combined with CNN as presented in **K5**. It is better to train word embedding on the dataset itself rather than using a pre-trained model especially when the domain is far away from the original domain of the pre-trained model.

In **K10** and **K21**, the accuracies of the DL models are the same. However, when the pre-trained word embedding model is applied in **K6**, we noted a significant difference in accuracies of the DL models.

a: EARLY DETECTION OF RUMORS

An automated debunking of false news at the stage of diffusion is named early rumor detection. Identification of trending rumors requires efficient models that can capture large-range dependencies among posts and produce distinct representations for accurate detection. This is difficult to handle using traditional ML approaches since these models are trained based on specific-event features, making them fail when transferred to unseen events Wang *et al.* [59]).

Recently, several early-rumor-detection models using DL have been proposed:

— In **K8,** Chen *et al.* [11] proposed an RNN-based deep attention model called *CallAtRumor*. This model learns the temporal hidden representation of sequential posts for rumor identification. CallAtRumor attempts to find distinct features by learning latent representation from sequential tweets. Then, it produces a hidden representation that captures the contextual variation of relevant tweets over time. CallAtRumor is also capable of dealing with duplicate data. The results of this study are compared with five state-of-the-art rumor detection approaches to illustrate that the proposed approach is sensitive to distinguishable words.

— In**K10,** Wang *et al.* [59] proposed an EANN framework for new rumors' detection that can identify false information based on multi-modal features. CNN was used to extract textual features, while a pre-trained VGG19 neural network was used to efficiently extract visual features (see Table 10). The proposed model learned invariant-event features so that it can detect newly emerged events. To achieve this goal, the authors eliminated the distinctive features of each event by measuring the dissimilarities between the different feature representations among different events.

— Similarly, in **K12,** Wu *et al.* [62] built a hybrid model to detect false information. Their proposed method uses the RCNN model, which is a CNN based model that uses a recurrent structure as a convolutional layer of the CNN model. The model uses a recurrent structure to capture semantic-text features and convNet model to learn sentiment features. Additionally, a deep bidirectional gated recurrent unit (D-Bi-GRU) is presented in **K7** to automatically detect rumor. The presented method is based on a feature-selection process to

TABLE 10. Features used in early rumors' detection studies.

Study	Features	Study	Features
K7, K8, K14, K15	Textual features	K10	Textual and visual features
K13	Content, user, and diffusion features	K12	Text semantic features and sentiment features
K16	Textural and structural features	K5	Contents, spreaders and diffusion features
K20	Local and global features of phrases as well as temporal tweet semantics	K24	Textual and temporal features

classify rumors and uses forward and backward user response sequences

— In **K16**, Ma *et al.* [42]focus on learning discriminative features from the non-sequential propagation structure of the tweets to generate a powerful representation allowing them to identify rumors. They proposed a two-recursive neural model based on top-down and bottom-up tree-structured neural networks to detect and classify rumors.

b: MAIN FINDINGS

— 4 out of 25 studies proposed models for the detection of multi-modal fake news. Pre-trained VGG19 neural networks are used to extract visual features in **K10**. In **K11**, the authors proposed a hybrid model of LSTM and CNN. First, the model extracts the features from the tweets without any prior information about the topic. Then, it classifies the rumor using conversations and images. The results of the proposed framework show an accuracy close to 82%. Authors in **K18** and **K19** use RNN with an attention mechanism (att-RNN) to fuse multimodal features.

— In **K9,** authors argued that stance and rumor detection should be treated jointly considering the strong connection between the veracity of the claim and the stances expressed in the responsive posts. The authors employ RNN in their proposed framework. Similarly, in **K15**, Poddar *et al.* (2018) proposed a two-step model. The first step aims to detect the stance of each tweet by considering time stamp and sequential conversation structure. The second step predicts the stances of all tweets in a conversation tree to determine the veracity of the original rumor.

— In **K6**, Roy *et al.* [57] proposed an ensemble-based architecture for fake news detection. They developed a model based on CNN and Bi-directional LSTM networks (BI-LSTM) to classify rumors. Results of CNN and BI-LSTM are then transferred to the multi-layered

---

[11]The exact numbers are not precise, since the authors use a figure

perceptron model to obtain the final classification. This model achieves an overall accuracy of 44.87%.

— Past research works have been unable to distinguish original posts and retweets for rumor detection. In **K13**, Xu et al. (2018) proposed a model named MNRD that considers three aspects to detect rumors, namely, original post content, diffusion of retweets, and user information. This model uses an attention mechanism to extract informative words and retweets in the diffusion process.

— In **K14**, Qian et al. proposed a novel two-level convolutional neural network with a user response generator to develop an automated model to capture semantic information from the text at word and sentence levels. In order to generate a response to new articles, the authors proposed a generative model that is learned from the historical user responses.

— In **K17**, a case study is presented to investigate the influence of the rumor ratio in the training dataset on the accuracy and performance of RNN based models. The rumor ratio in the training data must not be too low since a lack of rumors in the training data leads to a significant decrease in the performance of models.

— In **K24**, a method for rumor detection by learning the continuous representations of posts is detailed. The authors developed an RNN-based model to learn the hidden representations that capture the variation of contextual data of relevant posts between consecutive timestamps.

VI. CHALLENGES AND OPEN ISSUES

This section addresses the research questions *RQ6* and *RQ7*. The first question is related to the main challenges of DL with rumors' detection whereas the second question highlights open issues of DL on rumor detection field.

A. RUMOR DETECTION CHALLENGE

Many challenges face academic researchers when studying and analyzing the content of social media networks. Due to the accessibility of data, online citizens used Twitter as a primary space to publicly express their reactions to news/events Williams *et al.* [58]. Accordingly, Twitter has become the most preferred space where researchers can obtain data for research purposes. To address research question *RQ6*, we list some challenges next that are faced by the researchers when using social media data.

- **Ethical issues**: Twitter provides developers as well as the scientific community with a set of Twitter streaming Application Programming Interface (APIs), which make the data more accessible to researchers. However, collecting and retrieving data without informed consent produces several ethical issues, especially when the research deals with tweets on sensitive topics.

- **Legal issues**: Twitter needs some requirements for producing and sharing tweets within a publication. In most cases,

sharing datasets is prohibited since it requires informed consent from Twitter as well as from all of the participants.

- **Cost issue:** Twitter provides three levels of data access Williams *et al.*: (1) a free random 1% with ∼5M tweets daily; (2) chargeable or free 10%; and (3) 100% to academic researchers upon request and approval. Twitter provides data within 7 days according to the topic of interest. Thus, if the research requires obtaining more historical data, researchers can purchase data at different costs depending on the query and time of retrieval.

In addition to the aforementioned issues, there are some challenges regarding the dataset itself. Since building a reasonable dataset for detecting rumors through social media is a difficult task, we list below set of issues that researchers can face:

- **Dataset retrieval from the data source**: using hashtags and keywords to obtain tweets related to a specific topic does not guarantee the retrieval of all related data. In addition, using different keywords/hashtags to retrieve different data requires additional preparation and transformation of the resulting data such as conversion to the appropriate format and filtering non-relevant data after data retrieval.

- **Dataset bias:** using keywords-based queries and hashtags allows obtaining data written only in a specific language. Thus, datasets are also likely to be limited by the language that is used to retrieve data.

- **Representativeness:** user posts are not representative and the amount of posts depends mainly on the number of users who participate on social media. It can be the case that a rumor that spreads throughout a nation's offline population, but its impact on online social media is negligible due to the limited number of online users or due to political restriction.

- **Dataset quality issue:** the quality of a dataset is mainly affected by the dataset annotator (person who assesses the label classes of a dataset) or by enhancing the content of data. High dataset quality can be obtained by using two ways: (i) calculating inter-rater agreement where the κ statistic[12] can be a helpful technique (Eugenio and Glass [18]; Carletta *et al.* [10] and (ii) examining properties of text and authors, where the goal is to detect wither the content is recognized as spam or not Agichtein *et al.* [2].

B. FUTURE PATH AND DIRECTIONS

DL is a promising technology that can be used for rumor detection; however, researchers should be careful in selecting the appropriate DL architecture. DL models often require large datasets, which are not publicly available or need special permission from social media providers. Additionally, for training the DL, in many cases, there is not a clear methodology to understand how many layers should to used and which architecture is appropriate for the social analysis task. According to Zubiaga *et al.* [74], there are four types of rumor

---

[12]High κ statistic value denotes that the two annotators can reach a high level of agreement in identifying rumors (Yang *et al.* [69]

classification tasks, namely, *rumor detection*, *rumor tracking*, *stance detection*, and *veracity*. The answer to the research question **RQ7** is reported through the following recommendations:

— When working with small datasets, the RNN technique is not recommended for the early detection task as it has a bias towards the latest elements of the input sequence Cao *et al.* [8].

— DL techniques provide excellent results for not only rumor detection and tracking, but also for classifying rumor stance and veracity.

— Stand on individual social-media posts without looking at the use of context and interactions reduces the accuracy of rumor detection. Thus, further information extracted from user metadata can help to boost the performance of classifiers. We think that the combination of other NLP tasks can be useful.

— Most of the research studies use only textual features. We think that it is worthwhile to employ DL when dealing with visual formats such as videos and images.

— Since there are a few publicly available datasets, it would be valuable for the research community if more researchers share their datasets. This, in particular, will enable researchers to perform further studies over different datasets and compare the performances of their works.

— The researchers are encouraged to clearly specify tools used in their work as well as the performance obtained. This will enable the scientific community to reproduce the results and facilitate further extensions.

— As we stated earlier, the dataset is collected based on keywords used to collect posts associated with rumors that are known a priori. In most cases, retrieved data is written only in the language of the executed queries. We think that it is worthwhile to look for rumors written in different languages in future work.

VII. CONCLUSION

Deep learning (DL) has gained great success in many fields. Recently, several DL-based methods have been proposed for rumor detection on social networks. In this paper, a rigorous systematic literature review (SLR) about these methods has been conducted using 108 papers that were published from 2015 to 2019. At the final stage, 18 papers that emphasized the considered field were examined. Specifically, we started by detailing the main DL methods used in rumor detection. We provided a detailed comparison of the considered DL studies based on the method, performance, tool, and architecture. We also highlighted the principal datasets used by the research community according to the dataset's type, source, and content. Finally, we provided insights into the challenges and future directions for research on rumor detection using DL.

REFERENCES

[1] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Inf. Process. Manage.*, to be published.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. Int. Conf. Web Search Data Mining (WSDM)*, 2008, pp. 183–194.

[3] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on Twitter with hybrid CNN and RNN models," in *Proc. 9th Int. Conf. Social Media Soc.*, 2008, pp. 226–230. doi: 10.1145/3217804.3217917.

[4] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2015. doi: 10.1001/archdermatol.2012.2937.

[5] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, and Y. Kompatsiaris, "Verifying multimedia use at MediaEval," in *Proc. MediaEval*, 2015, pp. 1–3.

[6] A. J. Berinsky, "Rumors and health care reform: Experiments in political misinformation," *Brit. J. Political Sci.*, vol. 47, no. 2, pp. 241–262, 2017.

[7] G. Cai, H. Wu, and R. Lv, "Rumors detection in Chinese via crowd responses," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Piscataway, NJ, USA, Aug. 2014, pp. 912–917. [Online]. Available: http://dl.acm.org/citation.cfm?id=3191835.3192014

[8] J. Cao, J. Guo, X. Li, Z. Jin, H. Guo, and J. Li, "Automatic rumor detection on microblogs: A survey," 2018, *arXiv:1807.03505*. [Online]. Available: https://arxiv.org/abs/1807.03505

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.

[10] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Comput. Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[11] T. Chen, H. Chen, and X. Li, "Rumor detection via recurrent neural networks: A case study on adaptivity with varied data compositions," in *Trends and Applications in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science), vol. 11154, M. Ganji, L. Rashidi, B. Fung, and C. Wang, Eds. Cham, Switzerland: Springer, 2018.

[12] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Trends and Applications in Knowledge Discovery and Data Mining*, M. Ganji, L. Rashidi, B. C. M. Fung, and C. Wang, Eds. Cham, Switzerland: Springer, 2018, pp. 40–52.

[13] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[14] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[15] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, 2014.

[16] S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On deep learning for trust-aware recommendations in social networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1164–1177, May 2017.

[17] N. DiFonzo and P. Bordia, "Rumor, gossip and urban legends," *Diogenes*, vol. 54, no. 1, pp. 19–35, 2007. doi: 10.1177/0392192107073433.

[18] B. D. Eugenio and M. Glass, "The kappa statistic: A second look," *Comput. Linguistics*, vol. 30, no. 1, pp. 95–101, Mar. 2004.

[19] Y. Gao, X. Han, and B. Li, "A neural rumor detection framework by incorporating uncertainty attention on social media texts," in *Proc. Int. Conf. Cogn. Comput.* Cham, Switzerland: Springer, Jun. 2019, pp. 91–101.

[20] Y. Geng, Z. Lin, P. Fu, and W. Wang, "Rumor detection on social media: A multi-view model using self-attention mechanism," in *Proc. Int. Conf. Comput. Sci.* Cham, Switzerland: Springer, Jun. 2019, pp. 339–352.

[21] G. Grekousis, "Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis," *Comput., Environ. Urban Syst.*, vol. 74, pp. 244–256, Mar. 2019. doi: 10.1016/j.compenvurbsys.2018.10.008.

[22] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 943–951. doi: 10.1145/3269206.3271709.

[23] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-time credibility assessment of content on Twitter," in *Proc. 6th Int. Conf. Social Inform. (SocInfo)*, Barcelona, Spain, L. M. Aiello and D. McFarland, Eds. Cham, Switzerland: Springer, Nov. 2014, pp. 228–243. doi: 10.1007/978-3-319-13734-6_16.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[25] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.

[26] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 795–816. doi: 10.1145/3123266.3123454.

[27] Z. Jin, J. Cao, Y. Zhang, and Y. Zhang, "MCG-ICT at mediaEval 2015: Verifying multimedia use with a two-level classification model," in *Proc. MediaEval Multimedia Benchmark Workshop*, 2015, pp. 1–2.

[28] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. 13th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, Feb. 2016, pp. 2972–2978.

[29] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Keele, U.K., Durham Univ., Durham, U.K., Joint Rep. EBSE 2007-001, 2007. Accessed: Oct. 18, 2019. [Online]. Available: https://www.elsevier.com/data/promis_misc/525444systematicreviewsguide.pdf

[30] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882

[31] S. Kumar and N. Shah, "False information on Web and social media: A survey," 2018, *arXiv:1804.08559*. [Online]. Available: https://arxiv.org/abs/1804.08559

[32] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2013, pp. 1103–1108. doi: 10.1109/ICDM.2013.61.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015. doi: 10.1038/nature14539.

[34] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1096–1104.

[35] L. Li, G. Cai, and N. Chen, "A rumor events detection method based on deep bidirectional GRU neural network," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput. (ICIVC)*, 2018, pp. 755–759. doi: 10.1109/ICIVC.2018.8492819.

[36] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 99–108, Sep. 2015. doi: 10.1109/TCSS.2016.2517458.

[37] Y. Liu, X. Jin, and H. Shen, "Towards early identification of online rumors based on long short-term memory networks," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1457–1467, Jul. 2019. doi: 10.1016/j.ipm.2018.11.003.

[38] S. Lomborg and A. Bechmann, "Using APIs for data collection on social media," *Inf. Soc.*, vol. 30, no. 4, pp. 256–265, 2014.

[39] J. Ma, W. Gao, and K. F. Wong, "Detect rumors on Twitter by promoting information campaigns with generative adversarial learning," in *Proc. World Wide Web Conf.*, 2019, pp. 3049–3055.

[40] J. Ma, W. Gao, and K.-F. Wong, "Detect rumor and stance jointly by neural multi-task learning," in *Proc. Companion Web Conf.*, 2018, pp. 585–593. doi: 10.1145/3184558.3188729.

[41] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 1980–1989.

[42] 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 *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 3818–3824.

[43] J. Ma, W. Gao, and K. F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 708–717.

[44] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a large-scale corpus for evaluating event detection on Twitter," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 409–418.

[45] T. N. Nguyen, C. Li, and C. Niederée, "On early-stage debunking rumors on Twitter: Leveraging the wisdom of weak learners," in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, 2017, pp. 141–158.

[46] L. Poddar, W. Hsu, M. L. Lee, and S. Subramaniyam, "Predicting stances in Twitter conversations for detecting veracity of rumors: A neural approach," in *Proc. IEEE 30th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2018, pp. 65–72. doi: 10.1109/ICTAI.2018.00021.

[47] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, 2018, Art. no. 92.

[48] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3834–3840. [Online]. Available: http://dl.acm.org/citation.cfm?id=3304222.3304302

[49] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. ACM Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 797–806.

[50] (Feb. 16, 2019). *Rumor | Definition of Rumor in English by Oxford Dictionaries*. [Online]. Available: https://en.oxforddictionaries.com/definition/Rumor

[51] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. Artif. Intell. Statist.*, 2009, pp. 448–455.

[52] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015. doi: 10.1016/j.neunet.2014.09.003.

[53] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017. doi: 10.1145/3137597.3137600.

[54] C. Song, C. Tu, C. Yang, Z. Liu, and M. Sun, "CED: Credible early detection of social media rumors," 2015, *arXiv:1811.04175*. [Online]. Available: https://arxiv.org/abs/1811.04175

[55] C. R. Sunstein, *On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can Be Done*. Princeton, NJ, USA: Princeton Univ. Press, 2014.

[56] A. Roy, K. Basak, A. Ekbal, and P. Bhattacharyya, "A deep ensemble framework for fake news detection and classification," 2018, *arXiv:1811.04670*. [Online]. Available: https://arxiv.org/abs/1811.04670

[57] M. L. Williams, P. Burnap, and L. Sloan, "Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation," *Sociology*, vol. 51, no. 6, pp. 1149–1168, 2017.

[58] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2018, pp. 849–857. doi: 10.1145/3219819.3219903.

[59] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," 2017, *arXiv:1705.00648*. [Online]. Available: https://arxiv.org/abs/1705.00648

[60] J. Wehrmann, W. Becker, H. E. L. Cagnini, and R. C. Barros, "A character-based convolutional neural network for language-agnostic Twitter sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2384–2391.

[61] L. Wu, Y. Rao, H. Yu, Y. Wang, and A. Nazir, "False information detection on social media via a hybrid deep model," in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, Sep. 2018, pp. 323–333.

[62] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 651–662. doi: 10.1109/ICDE.2015.7113322.

[63] C. Wyrwoll, "Metadata in user-generated content," in *Social Media: Fundamentals, Models, and Ranking of User-Generated Content*. Wiesbaden, Germany: Springer, 2014, pp. 47–85. doi: 10.1007/978-3-658-06984-1_3.

[64] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[65] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 647–653. doi: 10.18653/v1/P17-2102.

[66] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[67] N. Xu, G. Chen, and W. Mao, "MNRD: A merged neural model for rumor detection in social media," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7. doi: 10.1109/IJCNN.2018.8489582.

[68] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proc. ACM SIGKDD Workshop Mining Data Semantics*, 2012, Art. no. 13.

[69] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3901–3907.

[70] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," 2018, *arXiv:1812.00315*. [Online]. Available: https://arxiv.org/abs/1812.00315

[71] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM neural network for text classification," 2015, *arXiv:1511.08630*. [Online]. Available: https://arxiv.org/abs/1511.08630

[72] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, 2017, Art. no. 32.

[73] A. Zubiaga, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media," 2016, *arXiv:1610.07363*. [Online]. Available: https://arxiv.org/abs/1610.07363

[74] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. K. Ward, "Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval," *CoRR*, vol. abs/1502.06922, 2015.

MOHAMMED AL-SAREM received the M.S. degree in information technology from the Faculty of Informatics and Computer Engineering, Volgograd State Technical University, Volgograd, Russia, and the Ph.D. degree from the Faculty of Informatics, University of Hassan II Casablanca, Mohammedia, Morocco, in 2007 and 2014, respectively. He is currently an Assistant Professor with the IS Department, Taibah University, Medina, Saudi Arabia. He has published several papers and participated in managing several international conferences. His current research interests include group decision making, multicriteria decision making, data mining, E-learning, natural language processing, and social analysis.

WADII BOULILA (SM'18) received the B.Eng. degree in computer science from the Aviation School of Borj El Amri, in 2005, the M.Sc. degree from the National School of Computer Science (ENSI), University of Manouba, Tunisia, in 2007, and the Ph.D. degree conjointly from ENSI and Telecom-Bretagne, University of Rennes 1, France, in 2012. He is currently an Assistant Professor of computer science with the IS Department, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia. His primary research interests include big data analytics, deep learning, data mining, artificial intelligence, uncertainty modeling, and remote sensing images. He is a Permanent Researcher with the RIADI Laboratory, University of Manouba, and an Associate Researcher with the ITI Department, University of Rennes 1. He has served as the chair, a Reviewer, and a TPC member for many leading international conferences and journals.

MUNA AL-HARBY received the B.Sc. degree in science of information systems from Taibah University, Saudi Arabia, where she is currently pursuing the master's degree in science of information systems. Her research interests include machine learning, text and web mining, and sentiment analysis.

JUNAID QADIR (M'14–SM'14) received the bachelor's degree in electrical engineering from the University of Engineering and Technology (UET), Lahore, in 2000, and the Ph.D. degree from the University of New South Wales, Australia, in 2008. He has been an Associate Professor with the Information Technology University (ITU), Lahore, Pakistan, since December 2015. He was an Assistant Professor with the School of Electrical Engineering and Computer Sciences, National University of Sciences and Technology, Islamabad, Pakistan, from 2008 to 2015. His primary research interests include computer systems and networking and using ICT for development (ICT4D). He has served on the program committee of a number of international conferences and reviews regularly for various high-quality journals. He has received the highest national teaching award in Pakistan—the higher education commissions (HEC) best university teacher award, from 2012 to 2013. He has considerable teaching experience and a wide portfolio of courses taught in the disciplines of systems and networking, signal processing, and wireless communications and networking. He is a member of ACM. He is an Associate Editor of IEEE Access, *Big Data Analytics Journal* (Springer), *Human-Centric Computing and Information Sciences* (Springer), and the *IEEE Communications Magazine*.

ABDULLAH ALSAEEDI received the B.Sc. degree in computer science from the College of Computer Science and Engineering, Taibah University, Medinah, Saudi Arabia, in 2008, the M.Sc. degree in advanced software engineering from the Department of Computer Science, The University of Sheffield, Sheffield, U.K., in 2011, and the Ph.D. degree in computer science from The University of Sheffield, U.K., in 2016. He is currently an Assistant Professor with the Computer Science Department, Taibah University. His research interests include software engineering, software model inference, grammar inference, machine learning, social network mining, data mining, and document processing.

● ● ●