The Future of False Information Detection on Social Media: New Perspectives and Trends

BIN GUO and YASAN DING, Northwestern Polytechnical University, P.R. China
LINA YAO, The University of New South Wales, Australia
YUNJI LIANG and ZHIWEN YU, Northwestern Polytechnical University, P.R. China

The massive spread of false information on social media has become a global risk, implicitly influencing public opinion and threatening social/political development. False information detection (FID) has thus become a surging research topic in recent years. As a promising and rapidly developing research field, we find that much effort has been paid to new research problems and approaches of FID. Therefore, it is necessary to give a comprehensive review of the new research trends of FID. We first give a brief review of the literature history of FID, based on which we present several new research challenges and techniques of it, including early detection, detection by multimodal data fusion, and explanatory detection. We further investigate the extraction and usage of various crowd intelligence in FID, which paves a promising way to tackle FID challenges. Finally, we give our views on the open issues and future research directions of FID, such as model adaptivity/generality to new events, embracing of novel machine learning models, aggregation of crowd wisdom, adversarial attack and defense in detection models, and so on.

CCS Concepts: • **Information systems** → **Social networks**; • **Human-centered computing** → *Collaborative and social computing*;

Additional Key Words and Phrases: False information detection, fake news, crowd intelligence, explanatory detection, social media

ACM Reference format:
Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The Future of False Information Detection on Social Media: New Perspectives and Trends. *ACM Comput. Surv.* 53, 4, Article 68 (July 2020), 36 pages.
https://doi.org/10.1145/3393880

This work was partially supported by the National Key R&D Program of China (2017YFB1001803) and the National Natural Science Foundation of China (No. 61772428, No. 61725205, No. 61902320, and No. 61972319).

Authors' addresses: B. Guo (corresponding-author), Y. Ding, Northwestern Polytechnical University, Xi'an, P.R. China; emails: guob@nwpu.edu.cn, yasanding@mail.nwpu.edu.cn; L. Yao, The University of New South Wales, Sydney, Australia; email: lina.yao@unsw.edu.au; Y. Liang and Z. Yu, Northwestern Polytechnical University, Xi'an, P.R. China; emails: {liangyunji, zhiwenyu}@nwpu.edu.cn.

0360-0300/2020/07-ART68 $15.00
https://doi.org/10.1145/3393880

1 INTRODUCTION

Social media platforms (such as Twitter,[1] Facebook,[2] and Sina Weibo[3]) have revolutionized the dissemination mode of information, which greatly improve the velocity, volume, and variety of information transmission. However, social media facilitates the rapid dissemination of both fact and false information. According to a recent survey by Knight Foundation,[4] Americans estimate that 65% of the news they see on social media is fake news. Besides, false information usually spreads faster, deeper, and wider in social networks [179].

The adversarial use of social media to spread misleading information poses a political threat [8]. For example, during the 2016 U.S. presidential election, as many as 529 different low-credibility statements were spread on Twitter [73] and approximately 19 million malicious bot accounts published or retweeted posts supporting Trump or Clinton,[5] which potentially influenced the election. In 2018, *Science* magazine published a theme issue about "Fake News," where they reported that fake statements can arouse people's feelings of fear and surprise [179], which contributes to social panic. For instance, a fake video named *Somalis 'pushed into shallow grave' Ethiopia* caused violent clashes between two races in Ethiopia[6] and a piece of online false information that suggested that onward travel restrictions had been lifted in Greece resulted in a Greek police clash with migrants.[7] The above examples show that the widespread false information poses a serious threat to the ecology of social information dissemination [91]. Social media users are exposed to plenty of messages on various topics every day. It is impractical and infeasible for users to judge each message credibility [140]. Therefore, it is urgent to detect false information on social media.

With the advent of the new media era, multimodal social media posts have gradually become mainstream on social media. Consequently, the future of online false information will extend beyond text to high-quality and manipulative information materials, such as images, videos, and audios on a massive scale with the rapid development of artificial intelligence (AI) [8]. For example, *DeepFakes* [44, 56] utilizes deep learning models to create audio and video of real people saying and doing things they never said or did, which makes false information ever more realistic and harder to discern. Though automatic false information detection is not a new phenomenon, it has been attracting increasingly much more public attention.

For facilitating the understanding and explaining of false information on web and social media, Kumar et al. [89] summarize and classify false information based on its intention and knowledge. According to intention, false information can be divided into *misinformation*, which refers to the false information created during an event evolution or the knowledge updating without the purpose to mislead [87, 150], and *disinformation*, which refers to the false information that misleads others intentionally for some purpose [36, 166]. According to knowledge, false information can be considered as *opinion based*, which expresses users' subjective opinions and describes some cases without a unique ground truth, and *fact based*, which is information that fabricates or contradicts an absolute ground truth [172]. In addition, there are some similar terms in the relevant literature, such as *rumors* and *fake news*. The term *rumor* usually refers to information that are not verified at the time of publication [204]. Therefore, rumors may turn out to be proved as true or false. Unlike rumors, the term *fake news* is widely used to refer the news articles that are intentionally and verifiably false [162]. We categorize these terms according to their intention, as shown in Figure 1.

[1] https://twitter.com/.

[2] https://www.facebook.com/.

[3] https://weibo.com/.

[4] https://www.poynter.org/ethics-trust/2018/americans-believe-two-thirds-of-news-on-social-media-is-misinformation/.

[5] https://firstmonday.org/ojs/index.php/fm/article/view/7090/5653.

[6] https://www.bbc.com/news/world-africa-46127868.

[7] https://www.bbc.com/news/world-europe-47826607.

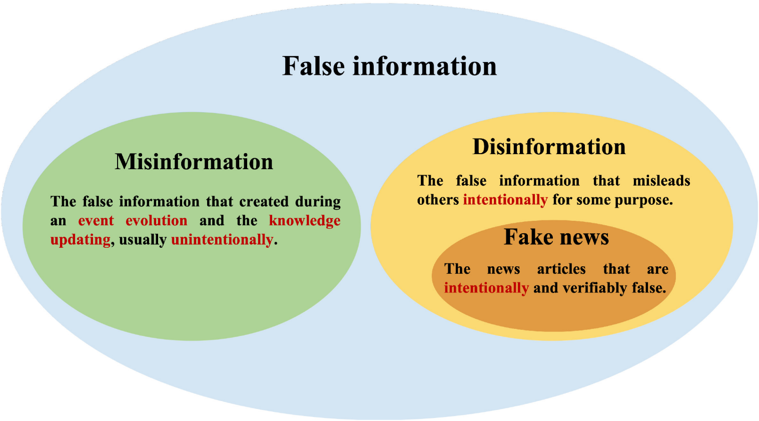

False information
Misinformation
The false information that created during an event evolution and the knowledge updating, usually unintentionally.
Disinformation
The false information that misleads others intentionally for some purpose.
Fake news
The news articles that are intentionally and verifiably false.

Fig. 1. The definitions of related terms and their relationships.

Although there are distinctions among the above terms, they all involve the dissemination of false information and have the ability or intention to affect some users. Therefore, this survey adheres the definitions of these terms and reviews the development of false information detection (FID) on social media from a technique perspective.

In recent years, there have been numerous efforts on FID. According to the type of features used in existing FID methods, we divide them into four categories: content-based methods, social context–based methods, feature fusion–based methods, and deep learning–based methods. The content-based detection methods mainly utilize textual or visual features extracted from social posts for binary classification (true or fake). The social context–based methods generally rely on the interaction characteristics among abundant users, such as commenting, reposting, and following. The feature fusion–based approaches make comprehensive use of content features and social context features. Moreover, deep learning–based methods mainly learn the latent depth representation of information through neural networks.

Although much research has been done on FID in the past few years, there are still numerous remaining issues to be addressed. First, existing FID methods mostly utilize content or propagation features and often work well on the entire lifecycle of false information, which may contribute to poor performance for early detection. Since false information could have a severe impact in just a few minutes,[8] it is crucial for detecting them at the early stage. Second, with the increase of multimodal posts propagating on social networks, traditional text-based detection approaches are no longer practicable, and it is beneficial to take advantage of images or videos for FID in more complex scenarios. Third, current detection methods only give the final result of whether the claim is fake but lacks reasons for the decision. It is significant to give a convincing explanation for debunking inaccurate information and preventing its further propagation.

This article aims to give an in-depth survey of the recent development related to FID methods. There have been several surveys on FID [39, 162, 201, 204]. Zhou et al. [201] study fake news from four perspectives, including knowledge based, style based, propagation based, and credibility based and summarize relevant detection methods in psychology and social science. Zubiaga et al. [204] focus on rumor classification systems and investigate the existing approaches for identifying suspected rumors, collecting rumor-related posts, detecting stances of posts, and evaluating the credibility of target events. Similarly, Fernandez et al. [39] divide the misinformation

[8] https://www.theverge.com/2013/4/23/4257392/ap-twitter-hacked-claims-explosions-white-house-president-injured.

detection into four phases: misinformation identification, propagation, validation, and refutation. They organize the existing online misinformation detection systems correspondingly. Shu et al. [162] divide detection models into news content-based models and social context–based models from the perspective of data mining and summarize the evaluation measurements of fake news detection algorithms. The differences between our survey and other related surveys are as follows:

- The above surveys pay little attention to deep learning–based false information detection methods. However, in the past three years, deep learning models have been widely used in FID. To provide an up-to-date comprehensive survey to detection methods, we investigate and cross-compare recent deep learning–based approaches.
- This article reviews the new issues and technologies emerging in the field of FID in recent years, such as *early detection*, *detection by multimodal data fusion*, and *explanatory detection*. Furthermore, our article surveys these new issues and promising works from the perspective of crowd intelligence, which investigates the potentials of leveraging crowd intelligence to facilitate FID.
- The development of AI has improved the performance of FID models, and thus datasets have become as important as algorithms. This article sorts out the widely used open datasets since 2015 for future researchers, which have been ignored by existing surveys.

Different from existing studies that mostly use the content of posts, crowd intelligence–based methods aim to detect false information based on aggregated user opinions, conjectures, and evidence, which are implicit knowledge injected during human–post interaction (e.g., publishing, commenting, and reposting of posts). Above all, the main contributions of our work include the following:

- Based on a brief literature review of FID, we concentrate on the recent research trends of it, including model generality to new events, early detection, multimodal fusion-based detection, and explanatory detection.
- We make an investigation of the crowd intelligence–based approach for FID, including the scope of crowd intelligence in FID, crowd intelligence–based detection models, and hybrid human–machine fusion models.
- We further discuss the open issues and promising research directions of FID, such as model adaptivity/generality to new events, embracing of novel machine learning models, and adversarial attack and defense in FID models.

The rest of this article is organized as follows. We give a brief literature review of existing FID works in Section 2. Then we investigate several new research trends in FID in Section 3. In Section 4, we highlight the crowd intelligence–based detection followed by open issues and future directions of FID in Section 5. Finally, we conclude this article in Section 6.

2 A BRIEF LITERATURE REVIEW

This survey primarily focuses on detecting false or inaccurate claims spreading on social networks, so we first give a general definition of the false information detection problem.

- For a specific statement s, it contains a set of related n posts $P = \{p_1, p_2, \ldots, p_n\}$ and a set of relevant m users $U = \{u_1, u_2, \ldots, u_m\}$. Each p_i consists of a series of attributes representing the post, including text, images, number of comments, and so on. Every u_i consists of a series of attributes describing the user, including name, register time, occupation, and so on.

- Let $E = \{e_1, e_2, \ldots, e_n\}$ refers to the engagements among m users and n posts. Each e_i is defined as $e_i = \{p_i, u_j, a, t\}$ representing that a user u_j interacts with the post p_i through action a (posting, reposting, or commenting) at time t.

Definition 2.1 (False Information Detection). Given a statement **s** with its posts set **P**, users set **U** and engagements set **E**, the false information detection task is to learn a prediction function $\mathcal{F}(\mathbf{s}) \rightarrow \{0, 1\}$, satisfying:

$$\mathcal{F}(s) = \begin{cases} 1, & \text{if } s \text{ is a piece of false information} \\ 0, & \text{otherwise} \end{cases}.$$

In the following, we give a brief literature review of existing FID techniques, categorized into four major types, namely *content based*, *social context–based*, *feature fusion–based*, and *deep learning–based* methods, as summarized in Table 1 (for the prior three types) and Table 2 (for the last type). Moreover, we also make a summary of several existing online FID tools, which are of significance to mitigate the impact of false information and prevent its further dissemination.

2.1 Content-based Methods

For a specific event, its microblog is generally composed of a piece of text to describe it, often associated with several pictures or videos. Content-based methods are mainly based on specific writing styles or sensational headlines in fake articles, such as lexical features, syntactic features and topic features [143]. For example, Castillo et al. [16, 17] find that highly credible tweets have more URLs, and the textual content length is usually longer than that of lower credibility tweets.

Many studies utilize the lexical and syntactic features to detect false information. For instance, Qazvinian et al. [136] find that the *part of speech (POS)* is a distinguishable feature for FID. Kwon et al. [90] find that some types of sentiments are apparent features of machine learning classifiers, including positive sentiments words (e.g., love, nice, sweet), negating words (e.g., no, not, never), cognitive action words (e.g., cause, know), and inferring action words (e.g., maybe, perhaps). Then they propose a periodic time-series model to identify key linguistic differences between true tweets and fake tweets. Moreover, Pérez-Rosas et al. [128] sum up the differences of linguistic features in real and fake contents, which can be divided into five categories: "Ngrams," "punctuation," "psycholinguistic features," "readability," and "syntax." A linear SVM is used to identify false information based on the above features. Rashkin et al. [141] summarize the language styles of untrusted news content. Specifically, they find that the first/second person pronouns are used more frequently in low-credibility information, and the same is true for exaggerated words.

Lexical features sometimes cannot fully reflect characteristics of false information because of its locality. Therefore, many studies introduce semantic features for FID, such as topic, sentiment, and writing style. For example, Potthast et al. [134] utilize different writing styles to detect fake claims. Similarly, Horne et al. propose an FID model based on the observation that fake news articles are substantially different from real news articles in their title style. Hu et al. [68] propose a framework for low-credibility social posts detection with sentiment information. Ito et al. [70] introduce the Latent Dirichlet Allocation (LDA) topic model into the evaluation of tweet credibility, and they propose tweet topic features and user topic features for detecting false information.

2.2 Social Context-based Methods

Traditional content-based methods analyze the credibility of the single microblog or claim in isolation, ignoring the high correlation between different tweets and events. Besides, a large amount of human–content interaction data (posting, commenting, reposting, rating and tagging, etc.)

Table 1. A Summary of Features Used by Existing Methods

Work	Feature Type	
	Content	**Social Context**
Castillo et al. [16]	*Containing question marks, sentiment, URL links, etc.*	*Propagation of initial tweets, max subtree, average degree, etc.*
Gazvinian et al. [136]	*Unigram, bigram, trigram, POS, hashtag, etc.*	*Propagation structure*
Hu et al. [68]	*Topic distribution, sentiment information*	———
Horne et al. [67]	*Language complexity, stylistic features*	———
Gupta et al. [61]	*First/second/third pronoun, exclamation marks, etc.*	*Follower-friend ratio, number of friends*
Shu et al. [164]	———	*Number of likes, number of followers, etc.*
Tacchini et al. [170]	———	*Number of likes*
Yang et al. [191]	———	*User opinions, viewpoints, user credibility*
Ma et al. [107]	*Topic distribution, question marks, exclamation marks, etc.*	*Average number of retweets and comments, etc.*
Liu et al. [102]	———	*User credibility, friendship*
Ma et al. [108]	———	*Syntactic parse tree and subtrees*
Jin et al. [74]	*Hashtag topic, URL links, etc.*	*Number of forwards and comments, propagation structure*
Della et al. [30]	*TF-IDF, stemmer*	*Number of likes*
Shu et al. [163]	*Content embedding*	*News-user social engagement embedding*
Kwon et al. [90]	*Positive words, negating words, cognitive action words, inferring action words, etc.*	*Clustering of friendship network, fraction of isolated nodes, etc.*
Wu et al. [186]	———	*Number of followers, number of comments and reposts, propagation tree*
Volkova et al. [176]	*Language complexity and readability, moral foundations, psycholinguistic cues, etc.*	*User opinions*
Ito et al. [70]	*tweet topics, user topics*	———
Pérez-Rosas et al. [128]	*Unigram, punctuation characters, word types, TF-IDF, etc.*	———
Long et al. [104]	*Topic distribution*	*User profiles (party affiliation, verified information, location)*
Rashkin et al. [141]	*First/second pronouns, strong subjective, modal adverbs*	———

provides abundant reference information for FID. Concretely, social context–based methods can be further divided into *post-based* and *propagation-based* methods.

(1) *Post-based features.* Post-based methods mainly rely on users' posts that express their emotions or opinions about specific events. Many studies detect false information by analyzing users' credibility [95, 118] or stances [63, 116]. For instance, Shu et al. [164] explore the truly useful features from user profiles for FID, so as to reduce the burden of feature extraction in the process of detection. Specifically, they find that extraverted and easygoing users are less likely to be affected by false information. Moreover, Guess et al. [57] state that conservatives are more inclined

Table 2. A Summary of Deep Learning–based Methods

Work	Model	Data Inputs			
		Text	Visual data	User response	User or website profiles
Ma et al. [106]	RNN	✓		✓	
Yu et al. [195]	CNN	✓		✓	
Jin et al. [72]	RNN	✓	✓	✓	
Li et al. [94]	GRU	✓		✓	
Liu et al. [99]	Attention	✓		✓	
Runchansky et al. [144]	RNN	✓		✓	✓
Chen et al. [20]	LSTM + Attention	✓		✓	
Nguyen et al. [123]	CNN + LSTM	✓		✓	
Guo et al. [60]	LSTM + Attention	✓		✓	✓
Popat et al. [133]	LSTM + Attention	✓			✓
Liu et al. [101]	CNN + GRU			✓	
Dong et al. [32]	GRU + Attention	✓			✓
Ma et al. [111]	GAN	✓		✓	
Yu et al. [194]	CNN + Attention	✓		✓	
Monti et al. [117]	GCN	✓		✓	✓

to share fake posts in Facebook. Long et al. [104] find that the application of user profiles (such as party affiliation, verified information, and location) in content-based detection methods can improve their performance on FID. They then propose a hybrid detection model, which extracts topic features of news content and user attribute features respectively. Furthermore, Tacchini et al. [170] find that social posts with inaccurate information usually have more likes than genuine facts. Therefore, they use a logistic regression (LR) model and a crowdsourcing algorithm to detect false information on the basis of users' likes.

(2) ***Propagation-based features.*** Propagation-based methods evaluate the credibility of posts and events as a whole [14], which usually pay attention to the construction of information dissemination network and credibility propagation.

Several studies detect false information by analyzing its propagation patterns. For instance, Ma et al. [107] find that features of social context gradually change over time. Therefore, they propose a *DSTS* model to characterize the temporal patterns of social context features for FID, which divides information propagation sequences into fixed length segments, then extracts both content-based and social context–based features from each segment of posts, and finally classifies them with SVM. Liu et al. [102] construct the information dissemination networks based on heterogeneous users' specific attributes for identifying special dissemination structures of false information. Kim et al. [79] propose a Bayesian nonparametric model to characterize the transmission of news articles, which jointly utilizes topics of articles and user interests for FID. Besides, Wu et al. [186] observed that fake messages are usually first published by an ordinary user, then forwarded by some opinion leaders, and finally spread by a large number of ordinary users. However, the truth is often posted by some opinion leaders and then directly spread by a large number of users. Based on this observation, they propose a hybrid SVM classifier for FID, which jointly models the message propagation structure, topical information, user attribute, and so on.

Besides, many studies also detect false information by constructing a specific tree or network structure. For example, Ma et al. [108] model the propagation of rumor-related microblogs as propagation trees, and they propose a kernel-based method to capture the patterns among those propagation trees for FID. Besides, Gupta et al. [62] construct a credibility propagation network containing users, posts and events to model the propagation process of false messages. Jin et al. [74] propose a three-layer credibility propagation network that connects microblogs, sub-events, and events for information credibility validation.

2.3 Feature Fusion–based Methods

Content-based detection methods mainly identify differences between true and untrue claims in terms of writing style and lexical and syntactic features, while social context–based detection methods mainly leverage features extracted from the process of information propagation. Since features applied by the two types of methods can be complementary [145], recently many researchers have begun to study novel feature fusion–based methods. For example, Vedova et al. [30] make use of the interactive information between users and posts, as well as the text information of posts. Specifically, they do stem analysis on social posts and represent each post as TF-IDF vector of words. Afterwards, they utilize users' *like* behaviors to describe social context features, similarly to the work of Tacchini et al. [170], and finally identify false information by integrating these two kinds of signals. To utilize traditional content features (e.g., lexical or syntactic features), Volkova et al. [176] use psycho-linguistic signals from the news content and authors' perspectives from social context as input data for different classifiers in FID. Besides, Shu et al. [165] further explore the social relations among publishers, news pieces, and users. They propose a universal detection framework called *TriFN*, which models the inherent relationship among news content, social interactions, and news publishers by nonnegative matrix factorization (NMF) algorithms for identifying low-credibility information.

2.4 Deep Learning–based Methods

Deep learning–based methods aim to abstract a high-level representation of false information data automatically. At present, most work mainly utilize Recurrent Neural Networks [106] and Convolutional Neural Networks [195] for FID, as presented in Table 2. In the following, we firstly summarize the widely used deep learning models, mainly including:

- *Convolutional Neural Networks (CNN).* CNN is one of the typical feedforward neural networks with three kinds of layers, i.e., convolutional layer, pooling layer, and fully connected layer [135]. In the convolutional layer, multiple filters (kernels) convolute with input vectors to generate feature maps. After that, the pooling layer reduces the dimension of feature maps to accelerate the training process of networks. Through multiple convolution and pooling operations, CNN can capture both local and global features from inputs. Finally, CNN outputs classification results by the fully connected layer (such as Softmax). It can be seen that FID models can capture content features between words and words, phrases and phrases by adjusting the size of filters.
- *Graph Convolutional Network (GCN).* GCN is a kind of neural network dealing with graph data, consisting of convolutional layers and fully connected layers, which can effectively capture structurally features of graphs [29, 83]. The hidden state matrix of each convolutional layer is obtained by a nonlinear variation of a special matrix, which is the product of the adjacent matrix of this graph and the hidden state matrix and weight matrix of its previous layer.
- *Recurrent Neural Networks (RNN).* RNN can effectively capture the features of sequential data, which saves former computations by information transmission between neurons in the same hidden layer. Social networking posts obviously have temporal characteristics, so

FID models can divide the interactive data of posts into continuous segments, and capture their sequential features by RNN. However, Glorot et al. [49] find that RNN may suffer from the gradient vanishing, which makes it not have a long-term memory. Therefore, long short-term memory (LSTM) [65] and gated recurrent unit (GRU) [25], a kind of RNN with gating mechanism, are widely used in NLP. LSTM adds a memory cell to store the current network state, and then controls the information flow through the coordination of the *input gate*, *forget gate* and *output gate*. Although GRU does not introduce additional memory units, it can control the current memory through a *reset gate* and an *update gate*.

- *Recursive Neural Network (RvNN).* RvNN is similar to RNN in that it unfolds the data structurally, which can be used to analyze hierarchical structures of data [135], such as the syntax analysis tree. This model is composed of a root node, left leaf nodes, and right leaf nodes. Besides, each node learns its representation from the direct left and right child nodes, which is calculated recursively until all nodes are traversed.
- *Auto-Encoder (AE).* AE is an unsupervised learning model including encoding and decoding stages [64]. In the encoding stage, the input data are transformed into latent vectors through multiple hidden layers, which will be reconstructed into the original data in the decoding stage. By minimizing the reconstruction error, AE learns the representation of inputs as much as possible. Compared with AE, variational auto-encoder (VAE) constrains the encoding stage and becomes a generative model [82]. Hidden layers of the encoding stage learn the latent variables by sampling from specific distributions, e.g., Gaussian distribution, and then input them to the decoding stage to generate realistic samples.
- *Generative Adversarial Network (GAN).* GAN is a kind of generative neural networks, consisting of a generator and a discriminator [51]. In the iterative process of backpropagation, the discriminator distinguishes whether its input comes from real datasets or fake samples generated by the generator, while the generator generates realistic samples based on the sampling distribution from datasets to confuse the discriminator. They finally achieve the Nash equilibrium, that is, the performance of the generator and discriminator cannot be improved any more.
- *Attention mechanisms.* Attention mechanisms are often used to describe the attention distribution of neural networks to input sequences [7]. It calculates the matching degree between current input sequences and output vectors, aiming at capturing key information of the inputs. The higher the matching degree, the higher the attention score. Accordingly, detection methods could utilize attention mechanisms to find these words or phrases that contribute more to FID.

Many existing studies utilize deep neural networks to learn latent textual representation of false information by modeling related posts as time-series data. For example, Ma et al. [106] propose a detection model based on RNN, which captures temporal-linguistic features of a continuous stream of user comments. Li et al. [94] consider that both the forward and backward sequence of post flows convey abundant interactive information, so they propose the Bidirectional GRU method for FID. Liu et al. [101] believe that there are differences between the propagation patterns of fake news and true news, and they utilize CNN and GRU to classify the propagation paths for identifying low-credibility information. Yu et al. [194] consider that time-series characteristics of posts contribute to modeling events accurately, and they propose the *ACAMI* model for FID. This model uses the event2vec (proposed to learn the event-related representation) and an attention mechanism to extract the temporal and semantic representation of events, and then uses CNN to extract high-level features for classifying fake microblog posts.

Some approaches combine textual information and social context information (such as user response, user or website profiles) as data inputs of deep neural networks. For instance, Guo et al. [60] propose a hierarchical neural network that considers information of users, posts, and propagation networks as data inputs. Besides, they leverage the attention mechanism to estimate distinct contributions of features in FID. The work of Ruchansky et al. [144] proposes a detection model based on RNN, which incorporates features of news content, user response, and the source users to promote the performance on FID. Ma et al. [111] propose a GAN-based detection model, which aims to capture the low-frequency but effective indications of fake tweets. The generator (a seq2seq model based on GRU) tries to generate controversial opinions, making the distribution of tweets' viewpoints more complicated, and the discriminator (based on RNN) attempts to identify robust features of false information from augmented samples.

There are also some works using graph neural networks for FID, such as GCNs. They often utilize neural networks to analyze the propagation structure of social posts, and then extract high-level representations of information propagation patterns for classifiers. For example, Monti et al. [117] propose a GCN-based FID model, which integrates the tweet content, propagation structure, user profiles, and user social relationships (following and being followed). Given the original tweet and all the related tweets, i.e., comments and retweets, the detection model takes each tweet as the node, tweet propagation paths and user relationships as the edge to build an event-specific graph. Afterwards, they use GCN to identify those low-credibility tweets, which contains two convolutional layers and two fully connected layers. In addition, Dong et al. [33] propose a GCN-based detection model, named *GCNSI*, which utilizes the graph convolutional networks to detect multiple sources of false information.

2.5 Existing Detection Tools

In addition to academic research, researchers have also developed several FID tools. According to their main detection content, existing online tools could be mainly divided into image-based and text-based tools.

- **Image-based detection tools**. *FotoForensics*[9] determines whether the target image has been modified by analyzing the distribution of image compression levels. If the image has been modified, then the claim containing this image is likely to be a piece of false information. Furthermore, the knowledge search engine, *Wolfram Alpha*,[10] also can verify the authenticity of images by retrieving information in its knowledge base, which contributes to detecting the false information containing images.
- **Text-based detection tools** [154, 160]. Existing FID tools mainly focus on the detection of textual content, and some tools have also become the basic reference for scholars to build datasets, such as *Politifact*,[11] *Snopes*,[12] and *Factcheck*.[13] They can check the facts of questionable claims circulating on social networks and provide users with analysis reports from trusted experts or journalists. In addition, researchers have also developed several online FID tools, such as *Fake News Detector*,[14] *TwitterTrails*,[15] and *Hoaxy*.[16]

[9] http://fotoforensics.com/.
[10] https://www.wolframalpha.com/.
[11] https://www.politifact.com/.
[12] https://www.snopes.com/fact-check/.
[13] https://www.factcheck.org/.
[14] https://fakenewsdetector.org/en.
[15] http://twittertrails.com/.
[16] https://hoaxy.iuni.iu.edu/.

3 NEW TRENDS IN FALSE INFORMATION DETECTION

Having reviewed the traditional studies on FID, this section investigates several new research trends of this field, including *early detection*, *detection by multimodal data fusion*, and *explanatory detection*.

3.1 Early Detection

False information can be readily spread by massive users on social networks, resulting in serious effects in a very short period [14, 46]. Therefore, early detection of false information becomes an important research topic. However, most existing studies (content-based and social context–based methods) detect false information by assuming that they have all the lifecycle data. They rely on several aggregation features, such as content characteristics and propagation patterns, which require a certain number of posts for training robust classifiers. The available data at the beginning of false information is so limited that it is challenging to detect it at the early stage. Recently, there have been some efforts for the early FID.

Traditional machine learning methods often analyze the user interaction information in the early propagation of posts, extract a large number of features manually, and finally use classifiers (e.g., SVM, Random Forest) to evaluate the credibility of them. For example, Liu et al. [100] find that source reliability, user diversity, and evidence signals, such as "I see" and "I hear," have significant influence on FID in a small amount of data. Besides, Qazvinian et al. [136] observe that users tend to express their own beliefs (e.g., supporting or questioning) in the early stage of tweets propagation. Therefore, the rational use of user beliefs in messages is of great benefit to the early detection of false information. To tackle the problem of lack of data, it will be another useful method to borrow knowledge from related events for FID. For example, Sampson et al. [149] propose a method for emergent FID by leveraging implicit linkages (e.g., hashtag linkages, web linkages) for additional information from related events. The experimental results show that such implicit links notably contribute to identifying emergent untrue claims correctly when less textual or interactive data is available.

Many detection approaches leverage deep learning models for early detection of false information. Deep learning–based detection methods often use neural networks to automatically extract social context features, and find key features of FID by utilizing attention mechanisms. For example, Liu et al. [99] observe that only a small number of posts contribute a lot to FID. To select these crucial contents, they propose an attention-based detection model, which evaluates the importance of each post by their attention values. Besides, the experiment results indicate that the proper usage of attention mechanism facilitates the early detection of false information. Similarly, Chen et al. [20] find that users tend to comment on different contents (e.g., from surprising to questioning) in different periods of information dissemination. Based on this observation, a deep attention model based on RNN is proposed to learn selectively temporal hidden representations of sequential posts for early FID. Yu et al. [195] utilize a CNN-based model to extract key features from sequences of posts and learn high-level interactions among them, which is beneficial for identifying fake tweets with relatively fewer interactive data. Nguyen et al. [123] also leverage CNN to learn latent representations of each tweet, obtaining the credibility of tweets accordingly. They then evaluate whether the target event is a piece of false information by aggregating all the predictions of related tweets at very beginning of the event. What's more, Liu et al. [101] find that most users retweet the source tweet without comments in the very early process of message propagation, implicitly causing some delay in utilizing user comments for early FID. Therefore, they propose a propagation path classification model, named *PPC*, which jointly uses CNN and GRU to extract local and global characteristics of users in retweeting paths.

Table 3. The Works of Multimodal FID on Social Media

Type	Work	Detection model
Low-level feature-based	Jin et al. [77]	SVM, LR, Random Forest
	Ferrara et al. [41]	GMM
	Salloum et al. [147]	CNN
	Huh et al. [69]	CNN+Siamese Networks
	Korshunov et al. [85]	SVM,PCA+LDA
	McCloskey et al. [112]	SVM
	Nataraj et al. [119]	CNN
High-level feature-based	Gupta et al. [61]	Naive Bayes,Decision Tree
	Jin et al. [75]	CNN
	Jin et al. [72]	Attention+LSTM+CNN
	Wang et al. [182]	Text-CNN+CNN
	Sabir et al. [146]	CNN+Word2vec
	Qi et al. [137]	CNN+GRU
	Khattat et al. [78]	VAE

3.2 Detection by Multimodal Data Fusion

Traditional FID methods focus on textual content and propagation structures. However, social media posts also contain rich visual data, such as images and videos, while such multimodal data is often ignored. Images and videos are more appealing to users than pure textual information, because they can vividly describe target events.

The great advances in image processing, such as AE, VAE, and GAN (as described in Section 2.4), have proved that images can be easily edited and modified, making fake images generation more readily. Consequently, analyzing the relationships among multimodal data and developing fusion-based models can be a promising way to FID [14]. There are mainly three kinds of fake images in false information on social media, including *image tampering*, *image mismatching*, and *image mixing*.

- **Image tampering**, malicious editing and modifying of existing images.
- **Image mismatching**, meaning that the image itself is real, but its text misinterprets it.
- **Image mixing**, using images of previous messages as the visual information of current messages.

When detecting multimodal false information, we could not know which type of fake images are included in social media posts in advance, so FID models need to extract pervasive features for effective detection. Existing works of multimodal FID are mainly divided into low-level feature-based methods and high-level feature-based methods, as presented in Table 3.

(1) ***Low-level feature-based methods.*** The most straightforward way to evaluate the credibility of a multimodal post is to verify the authenticity of its visual information. Low-level feature-based methods mainly analyze the discrete cosine transform (DCT) coefficients, patterns of color filter array (CFA) interpolation, and other low-level features of images to determine whether input samples have been edited or tampered.

For example, Jin et al. [77] find that there are some differences in the statistical characteristics of visual information between true and fake posts. Specifically, a fake event often has a few pictures that are repeatedly transmitted (limited by the source of pictures), while pictures in facts often have a strong diversity. Therefore, they propose several visual and statistical features to describe the differences for FID, such as *visual coherence score* and *visual diversity score*. Based on the assumption that the tampered and non-tampered regions of the same image come from different cameras, Ferrara et al. [41] utilize the Gaussian mixture model (GMM) to analyze the statistical features of CFAs in different regions of input images for detecting fake images. Fridrich et al. [45] propose the spatial rich model, which detects fake images by capturing the discontinuity of noise patterns of adjacent pixels in tampered and non-tampered regions. Furthermore, Salloum et al. [147] put forward a multi-task detection model named *MFCN*. This model uses CNN to extract tampering features from the surface and boundary of samples, which can identify splicing and local removal issues in images. Huh et al. [69] propose a CNN-based fake image detection model, which uses the image exchangeable image file format (EXIF) metadata to determine whether the content of images can be generated by the same imaging pipeline.

Moreover, Korshunov et al. [85] investigate that existing *Facenet*-based face recognition algorithms [152] are vulnerable to fake images and videos generated by GAN. Therefore, the research on detecting multimodal social media posts with generated pictures/videos has become increasingly important. McCloskey et al. [112] find that GAN-generated images have more overlaps than real images in the spectral response of RGB channels. In addition, the spectral response functions of real images are generally non-negative values, while the fully generated images often do not have this constraint. Therefore, they propose the *intensity noise histogram* and *saturation* to identify fake images. In addition, Nataraj et al. [119] utilize co-occurrence matrices to characterize the spatial consistency of images, and then they use CNN to learn the patterns of GAN-generated images in co-occurrence matrices for FID.

(2) ***High-level feature-based methods.*** The multimodal posts contain not only the images that have been maliciously tampered, but also the real images that have been wrongly used to report unrelated events. Therefore, it is difficult to identify this kind of false information simply by using image low-level features. High-level feature-based methods mainly utilize the semantic features of images and texts for multimodal FID.

For instance, Jin et al. [75] find that images in low-credibility posts are often expressed as shocking content, such as violence and horror, and often contain strong emotions. Gupta et al. [61] analyze 10,350 tweets with fake images circulated during Hurricane Sandy in 2012. They conclude that FID could benefit from the temporal characteristics, influence patterns, and user responses produced in tweets sharing process. Furthermore, Jin et al. [72] propose an attention-based multimodal FID model, called *att-RNN*, which integrates textual feature representation (learnt by LSTM) and visual feature representation (learnt by CNN). In att-RNN, the attention mechanism is used to extract key factors from text content and social context, and attention degrees can adjust the weight of visual semantic information to guide CNN to extract event-related semantic features. Similarly, Wang et al. [182] propose a detection model based on adversarial learning, named *EANN*, which utilizes Text-CNN [81] to extract text modal features and CNN to extract visual modal features. Sabir et al. [146] present a deep multimodal model for FID, which simultaneously utilizes the CNN, word2vec [114], and global positioning system to extract unique features of fake pictures.

Tampered images often show periodic characteristics in frequency domain, which can be used as a basis for multimodal FID. Therefore, Qi et al. [137] propose the *MVNN* to extract the frequency domain features (based on CNN) and pixel domain features (based on CNN-GRU) of images. *MVNN* uses an attention mechanism to adjust the weight of frequency domain and pixel domain features, based on the observation that the low-level features and high-level features are complementary in multimodal FID, i.e., they play different roles in different social media posts. Khattar et al. [78] believe that the simple concatenation of text feature vectors and visual feature vectors is difficult to fully express the association between the two modal information, and they propose the *MVAE*, a variational autoencoder-based FID model. In the encoding stage, *MVAE* encodes the textual feature representation (learnt by LSTM) and visual feature representation (learnt by CNN) into a shared latent vector through a fully connect layer. Afterwards, the reconstruction loss and the K-L divergence between the sampling distribution and Gaussian distribution are used to ensure that the encoded latent vector can be decoded back to the original state. This detection method then uses the encoded latent representation to identify whether the multimodal post is fake.

3.3 Explanatory False Information Detection

Most deep learning–based FID methods often do not present the reason for making decisions when they output the decision results, which utilize pre-trained classifiers to identify suspected events in the test set [14]. However, finding pieces of evidence that support decisions would be beneficial in debunking the false information and preventing its further spreading. Consequently, explanatory FID has become another trending research topic. Existing explanatory FID studies mainly focus on two aspects: one is to explore practical interpretable detection models (interpretation of models), and the other is to explain their results (interpretation of results).

(1) *Interpretation of models.* The research on interpretable FID models mainly focuses on the utilization of probabilistic graph models (PGMs) and knowledge graphs (KGs):

- *Probabilistic Graph Model (PGM).* PGMs use graphs to represent the joint probability distribution of correlational variables (nodes), consisting of the *Bayesian networks*, which use directed acyclic graphs to model the causal relationship between variables, and *Markov networks*, which use undirected graphs to model the interaction between variables [38]. The relationship of nodes can be explained by conditional independencies. Probabilistic graph-based detection models could simultaneously characterize the users, social posts and human–content intractions, and infer the implicit information credibility approximately according to the explicit interaction data. In addition, the widely used approximate inference algorithms are *variational inference*, *belief propagation*, and *Monte Carlo sampling* [84].
- *Knowledge Graph (KG).* KGs describe entities in the real world and the relationships among them in the form of graph. Specifically, KGs contain knowledge of various domains, define possible categories and relationships of entities, and allow any entities to be potentially associated with each other [127]. Furthermore, there are several authoritative knowledge bases, such as *Freebase*,[17] *Wikidata*,[18] *DBpedia*,[19] *Google's Knowledge Graph*.[20] These knowledge bases contain millions of entities and statements, which provide a reference for FID. Detection approaches could check the fact of social media posts by knowledge extraction, fusion and completion.

Specifically, Shi et al. [157] propose a KG-based fact-checking method. It firstly analyzes the semantic information of posts by extracting meta paths of similar entities from a knowledge graph. Afterwards, this method mines heterogeneous connection patterns in collected fact statements for fact-checking. Moreover, Gad-Elrab et al. [47] propose the *ExFaKT* to provide human-understandable explanations for candidate facts, which combines the semantic evidence from text contents and knowledge graphs. *ExFaKT* uses Horn rules, a subset of first order predicate logic, to rewrite target facts as multiple easy-to-explain facts for further FID. The work of Popat et al. [131] proposes a probabilistic model to unite the content-aware and trend-aware evaluating

[17] http://www.freebase.com.

[18] https://www.wikidata.org/wiki/Wikidata:Main_Page.

[19] https://wiki.dbpedia.org/.

[20] https://developers.google.cn/knowledge-graph/.

algorithms for FID. Specifically, they model mutual interactions among event-related articles to generate appropriate user-interpretable explanations, including linguistic features, stances, and the reliability of sources. Yang et al. [191] propose an unsupervised FID method called *UFD*, which utilizes Bayesian networks to model the complete generation process of truths and user opinions. *UFD* considers the authenticity of news articles and users' reputation as latent variables, and then exploits social engagements among users to extract their viewpoints on news credibility.

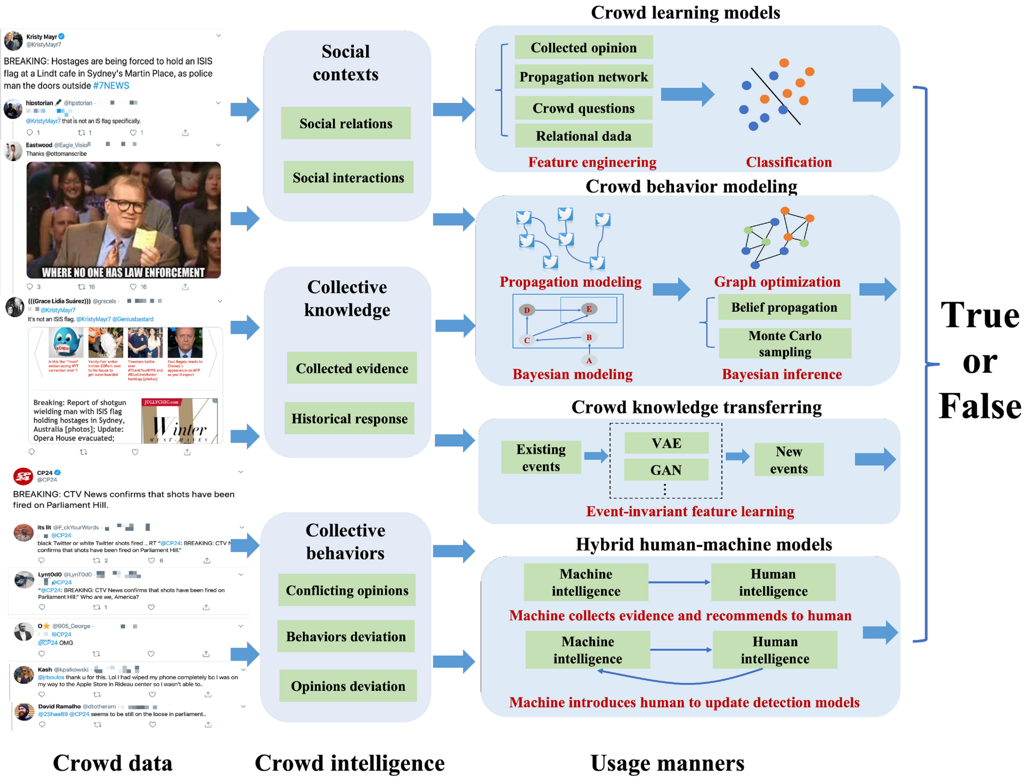
(2) *Interpretation of results.* The interpretation of results mainly refers to the visualization of decision-making process, or the analysis of facts. Although the deep learning–based method greatly improves the performance of FID, deep model is not well interpretable with its intrinsic mechanisms. Therefore, researchers utilize other auxiliary information for explanatory FID.

Since the attention degree in attention mechanisms could characterize the importance of each part of inputs [19], several deep learning–based detection methods explain their classifications through the visualization of attention degrees. For example, Chen et al. [20] visualize the attention distribution of some fake claims identified by their model, and find that most event-related words are given lower degrees than the words expressing users' doubt, anger and other emotions. Dong et al. [32] propose an attention-based FID model named *DUAL*, which uses GRU to extract textual features and DNN to extract social context features respectively. They visualize the attention matrices of two hidden layers, effectively depicting the distribution of attention degrees of each hidden layer when identifying true and fake posts. Similarly, Popat et al. [133] use a bidirectional LSTM to extract features of the source claim and external related posts respectively, and then combine the attention mechanism to learn the representation of false information. They also visualize its attention degrees, showing that many signal words such as "barely true," "evidence," and "reveal" are given higher degrees. In addition, they use the principal component analysis (PCA) to visualize the text feature vectors learned by their model, and find that the textual representation of true and false claims can be properly separated.

ClaimVerif [200] is an online explanatory information credibility evaluation system, which takes stance, viewpoint, source credibility and other factors of the given claim into account for providing valid evidence. When identifying fake claims online, *ClaimVerif* uses google search to crawl relevant articles, analyzes the textual features of the original message and repost messages, and finally outputs the credibility of the source message, along with human-understandable evidence. Similarly, *CredEye* [132] determines whether a given claim is fake by analyzing online related articles. The explanation is based on the language style, stance of these articles, and source reputations. Moreover, Yang et al. [190] present an explainable FID framework, named *XFake*, which comprehensively analyzes the attributes (such as subject, speaker, and context), semantic features, and linguistic features of statements. *XFake* displays several supporting examples through a visual interface, and shows the reasoning process in the form of ensemble trees.

4 CROWD INTELLIGENCE–BASED DETECTION

Existing studies show that content features of posts are still the top priority of FID. As social posts are generated, interacted, and consumed by users, it will intake various *human intelligence* (e.g., opinion, stance, questioning, evidence provision) in the editing, commenting, and reposting of posts. The so-called *crowd intelligence* [58, 96, 185] is also aggregated at a collective manner during the dissemination process of a social media post. As stated by Castillo et al. [16], a promising hypothesis is that there are some intrinsic signals in the social media environment that contribute to assessing the credibility of information. Ma et al. [110] also find that Twitter supports "*self-detect*" of false information based on aggregated user opinions, conjectures, and pieces of evidence. Though, how to leverage crowd intelligence in FID is still an open problem. In Section 4, we attempt



Kristy Mayr @KristyMayr7
BREAKING: Hostages are being forced to hold an ISIS flag at a Lindt cafe in Sydney's Martin Place, as police man the doors outside #7NEWS

hipstorian @hpstorian · [redacted]
@KristyMayr7 that is not an IS flag specifically.

Eastwood @Eagle_Visio [redacted]
Thanks @ottomanscribe

(((Grace Lidia Suárez))) @gracels · [redacted]
@KristyMayr7
It's not an ISIS flag. @KristyMayr7 @Geniusbastard

Breaking: Report of shotgun wielding man with ISIS flag holding hostages in Sydney, Australia [photos]; Update: Opera House evacuated;

CP24 @CP24
BREAKING: CTV News confirms that shots have been fired on Parliament Hill.

its lit @F_ckYourWords · [redacted]
@CP24
black Twitter or white Twitter shots fired .. RT "@CP24: BREAKING: CTV N confirms that shots have been fired on Parliament Hill."

LyntOdO @LynTOdO · [redacted]
@CP24
"@CP24: BREAKING: CTV News confirms that shots have been fired on Parliament Hill." Who are we, America?

O★ @905_George · [redacted]
@CP24
@CP24 OMG

Kash @kpalkowski · [redacted]
@jrboulos thank u for this. Lol I had wiped my phone completely bc I was on my way to the Apple Store in Rideau center so I wasn't able to..

David Ramalho @dtotheram · [redacted]
@2Shae89 @CP24 seems to be still on the loose in parliament..

Fig. 2. A taxonomy graph of crowd intelligence–based methods.

to address this problem by distilling and presenting several different forms of usage of crowd intelligence in FID systems, as shown in Figure 2.

4.1 Crowd Intelligence in False Information

In FID, crowd intelligence refers to aggregated cues or social signals from the wisdom of social media users during the information generation and dissemination process. In this subsection, we summarize the meaning and usage manners of crowd intelligence in FID.

We describe crowd intelligence from three aspects, including *social contexts*, *collective knowledge*, and *collective behaviors*.

- ***Social contexts.*** The social relations and interactions among source users and disseminators are helpful to understand the certainty of information. For example, Kim et al. [80] believe that the user flagging could indirectly reflect the credibility of tweets, so they use PGMs to generate the human–content interactive process and infer the truth of tweets. Zhao et al. [199] find that crowd questions or enquiries about the veracity in their comments are indicative signals of low-credibility information, and they use regular expressions to extract above signals from user comments for FID. Furthermore, Wu et al. [188] consider that similar topics may spread among similar crowd, so they encode the disseminators to capture their social proximity for identifying fake messages.
- ***Collective knowledge.*** The collected evidence provided by the crowd are useful to infer the credibility of information. For example, Lim et al. [97] utilize users' support or opposition

to the online evidence of target events to detect inaccurate statements. Rayana et al. [142] consider that users' ratings and comments are true evaluation of the credibility of posts, so they propose a detection framework named *SpEagle*, which extracts features from collective clues and relational data (information dissemination network). In addition, Qian et al. [138] propose a crowd knowledge transfer method for FID, where the crowd response knowledge (e.g., contextual features and behavioral features) from historical true/fake claims are leveraged.

- ***Collective behaviors.*** In many cases, though individual behaviors cannot well characterize the information credibility, the aggregated behaviors from a group of users often reveal more information. This may refer to crowd interaction patterns, behaviors or opinions deviation from the majority [88], conflicting viewpoints, and so on. For example, users who often participate in the production and dissemination of low-credibility information have behaviors deviation, e.g., posting several opinions in a short time, or interacting with contents after a fixed interval. Based on the above observation, Kumar et al. [88] infer the credibility of repliers and their comments by Bayesian modeling. Moreover, Jin et al. [76] find that related tweets under the same event contain supporting and opposing opinions (analyzed by the LDA topic model), and they utilize these conflicting viewpoints to build a credibility propagation network for FID.

Having investigated existing FID studies, we distill four different manners of usage of crowd intelligence, as presented below.

- ***Crowd learning models.*** It mainly uses feature engineering and representative learning to incorporate crowd intelligence in FID models.
- ***Crowd behavior modeling.*** It uses graph or probabilistic models to model crowd behaviors and interactions to infer the credibility of information.
- ***Crowd knowledge transferring.*** The learned FID models usually do not work well on new events. This manner tackles how to transfer crowd knowledge from existing events to new events.
- ***Hybrid human–machine models.*** Considering the complementary nature of human intelligence and machine intelligence, this manner concentrates on developing hybrid human–machine models for FID.

One common character of the prior three manners is that crowd intelligence is used in an implicit manner, without explicit human inputs. Specifically, crowd intelligence is represented as statistical human behavior patterns, used as features or parameters in the learning model. The last manner, however, is based on explicit human inputs, such as using crowdsourcing for data labeling. Thereafter, we describe related work about the prior three forms in Section 4.2 and the last form in Section 4.3.

4.2 Implicit Crowd Intelligence Models

In this section, we present the pioneering studies on the usage of implicit crowd intelligence for FID, particularly focusing on the first three manners depicted in Section 4.1, as summarized in Table 4.

(1) ***Crowd learning models.*** In this model, crowd intelligence is represented as features to train classifiers for detecting false information. This has been proved useful for the early FID. For instance, Liu et al. [100] try to solve the problem of real-time fake claims debunking using crowd cues from Twitter data, including people's opinion, statistics of witness accounts, aggregated belief to the event, network propagation, and so on. Zhao et al. [199] observe that some people are willing

Table 4. The Usage of Crowd Intelligence in FID

Usage manner	Work	Problem tackled	Usage of crowd intelligence
Crowd learning models	Liu et al. [100]	Early detection	Features by collected opinion, belief, etc.
	Zhao et al. [199]	Early detection	Features by crowd questions or enquires about the veracity.
	Wu et al. [188]	General	Features by social relations and propagation network.
	Rayana et al. [142]	General	Features by collective opinion clues and relational data.
Crowd behavior modeling	Hooi et al. [66]	General	Bayesian modeling, behavior deviation.
	Kumar et al. [88]	General	Bayesian modeling, behavior deviation.
	Jin et al. [76]	Early detection	A credibility propagation network model that incorporates conflicting social viewpoints.
	Ma et al. [110]	Early detection	Modeling reply structures and opinions by tree-structured recursive neural networks.
Crowd knowledge transfer	Wang et al. [182]	Early detection & Multimodal data fusion	The *Event Adversarial Neural Network* model to derive event-invariant features.
	Qian et al. [138]	Early detection	A generative Conditional Variational Autoencoder to transfer user response knowledge.
	Wu et al. [187]	Early detection	A sparse representation model for shared feature learning.

to question or inquire about the veracity of claims in Twitter before deciding whether to believe this message. Particularly, they find that the usage of enquiring minds facilitates early detection of false information.

Social relations and interactions are also widely used crowd intelligence in FID feature learning. For instance, Wu et al. [188] assume that similar messages often conduce to similar information propagation traces. They propose a social media user embedding method to capture the features of social proximity and social network structures, atop which an LSTM model is utilized to classify the information propagation path and identify its veracity. Rayana et al. [142] apply collective opinion clues and relational data to detect false information.

It is also helpful to identify false information by leveraging the crowd intelligence that user behaviors of publishing fake posts diverge from those of publishing genuine facts. Chen et al. [22] propose an unsupervised learning model that combines RNNs and Autoencoders to distinguish low credibility information from other authentic claims. Besides, Xie et al. [189] observe that the review spam attacks are strongly correlated with their rating patterns, which are distinct from the normal reviewers' behavior patterns. Therefore, they propose a review spam detection method based on their temporal behavior patterns, which provides a reference for FID by crowd learning models.

(2) *Crowd behavior modeling.* In this model, collective crowd behaviors, one type of crowd intelligence, are modeled as graphs or probabilistic models to infer information credibility. Hooi et al. [66] discover that fraudulent accounts often present their ratings in short bursts of time (rating scores satisfying the skewed distribution). The crowd wisdom is characterized by a Bayesian inference model, which can estimate how much a user's behaviors deviate from those of the related community. They infer the credibility of user ratings by measuring the degree of behavioral biases. Similarly, Kumar et al. [88] propose a Bayesian detection model that incorporates the aggregated crowd wisdom, such as behavior properties of users, reliability of ratings, and goodness of products. By penalizing unusual behaviors, it can infer the message credibility in rating platforms.

Some studies leverage aggregated crowd behavior modeling to facilitate early detection of false information. For example, Ma et al. [110] assume that the repliers are inclined to enquiry who supports or denies the given event and express their desire for more evidence. They thus propose two tree-structured recursive neural networks (RvNN) for effective fake tweets representation learning and early detection, which can model user reply structures and learn to capture the aggregated signals for FID.

(3) ***Crowd knowledge transfer.*** Existing FID models still do not perform well on emerging and time-critical events. In other words, existing FID models usually capture many event-dependent features that are not common to other events. Therefore, it becomes necessary to learn and transfer the shared knowledge learned from existing crowdsourced data to new events. The work of Wang et al. [182] proposes a detection model for identifying newly generated fake events using transferable features, named Event Adversarial Neural Network (*EANN*), which comprises three parts, i.e., "feature extractor," "event discriminator," and "fake news detector." The *EANN* uses the *event discriminator* to learn event-independent sharing features, and reduces the influence of event-specific features during model training.

Crowd knowledge transfer models also facilitate early FID. For example, Qian et al. [138] propose a generative Conditional Variational Autoencoder to capture user response patterns from historical users' comments on true and fake news articles. In other words, crowd intelligence is leveraged to generate responses toward new articles to improve the detection capability of models when social interaction data are not available at early stage of false information propagation. Wu et al. [187] also explore whether the detection of emerging fake social media posts could be benefited by the knowledge from historical crowdsourced data. They observe that social posts with similar contents often leads to similar behavior patterns (e.g., curiosity, inquiry). Consequently, a sparse representation model is built to select shared features and train event-independent classifiers.

4.3 Hybrid Human–Machine Models

FID is a challenging problem, and merely automatic models cannot well adapt to various contexts and events. Human intelligence, however, can appropriately remedy this problem by leveraging their knowledge and experience. Hybrid human–machine models are thus developed to harness the complementary nature of human intelligence and machine intelligence for FID. Broadly speaking, it belongs to the "human computation" paradigm, which aims to develop human–machine systems that interweave crowd and machine capabilities seamlessly to accomplish tasks that neither can do along [113, 177]. There are several representative examples of human–machine systems. For example, *reCAPTCHA* [178] is a Captcha-like system to protect computer security, while at the same time it harnesses combined efforts of individuals to the digitization of books. *Pandora* [124] is a hybrid human–machine approach that cab explain failures in component-based machine learning systems.

Different from the implicit crowd intelligence–based models, the human intelligence used in such models are often layered on explicit human inputs. Such models often present sufficient interpretable information to facilitate human–computer collaboration. With the aid of human–machine collaboration, the proposed models generally speed up the detection of false information.

Several human–machine models have been built for fact-hecking, as summarized in Table 5. For instance, Nguyen et al. [120] consider that a reliable system should be transparent to users on how

Table 5. The Usage of Hybrid Human–Machine Models

Work	Problem tackled	Usage of crowd intelligence
Nguyen et al. [120]	Explanatory detection	The mixed-initiative approach to blend human and machine intelligence.
Nguyen et al. [121]	Explanatory detection	Incorporate explicit crowd intelligence in the probabilistic graphical model.
Vo et al. [175]	Early detection	The machine recommends evidence URLs to human guardians to facilitate fact-checking.
Kim et al. [80]	Early detection	Using the marked temporal point processes to model crowd flagging procedure.
Tachiatschek et al. [174]	Early detection	Using the Bayesian inference mode to incorporate crowd flagging behavior.
Lim et al. [97]	Explanatory detection & Early detection	An interactive framework where machines collect pieces of evidence from Web search and human can give feedback to the evidence.
Bhattacharjee et al. [11]	Early detection	An active learning model that introduces human–machine interaction to update the detection model.

to get decision results. They propose a mixed-initiative approach that blends human knowledge and experience with AI for fact-checking. Besides, Nguyen et al. [121] present a hybrid human–machine approach based on PGMs, which integrates explicit human intelligence (by crowdsourcing) with the computing power to jointly model stance, veracity, and crowdsourced labels. This approach is capable of generating interpretations for FID. Vo et al. [175] present a fact-checking URL recommendation model to stop people from sharing false information. This model motivates guardians (users who tend to correct false information) to actively participate in fact-checking activities and spread verified articles to social networks.

Explicit human intelligence is also characterized and used in probabilistic models for FID. The work of Kim et al. [80] proposes the *CURB*, which leverages marked temporal point processes to model crowd-powered flagging procedure for low-credibility articles. To significantly mitigate the propagation of false information with provable guarantees, *CURB* can decide which statement to choose for identifying and when to check it. Tschiatschek et al. [174] also present a Bayesian inference model that incorporates crowd flagging for detecting fake posts. To assess the credibility of new claims from tweets, Lim et al. [97] present an interactive framework called *iFACT*. It collects independent evidence from web search results and identifies the dependencies between the new claims and historical claims. Users are further allowed to provide explicit feedback on whether the web search results are relevant (supporting or opposing) to unverified information. Besides, Bhattacharjee et al. [11] propose a human–machine collaborative learning system for fast refutation of false information. In this work, they introduce the active learning method into FID, first training an initial classifier on the limited labeled data, then using an interactive method to gradually update this detection model.

5 OPEN ISSUES AND FUTURE DIRECTIONS

Though researchers have made increasingly considerable efforts to address above challenges in FID systems, there are still open issues to be studied in the future, as discussed below.

(1) ***Cognitive mechanisms of false information.*** The research on people's cognitive mechanism of false information has promising guidance for the detection and refutation of fake social media posts [87], especially for the crowd intelligence–based detection methods. Several works conduct analysis of low-credibility posts on social media platforms to study the reasons why false information can spread quickly and widely. Lewandowsky et al. [92] consider that combating false information requires scientific research in the context of technology and psychology, and thus they propose a cross-disciplinary solution called "technocognition". Furthermore, they divide users' cognitive problems in the face of false information into four categories, including *influence effect*, *familiarity backfire effect*, *overkill backfire effect*, and *worldview backfire effect*, which lay a foundation for the research on users' perception of false information [93]. As concluded by Acerbi [1], the rapid dissemination of inaccurate information lies in that they contain specific contents satisfying users' cognitive preferences. For exploring the cognitive characteristics of false information, they further analyze the distribution of preferences in real and fake news articles by encoding cognitive preferences into "threat," "disgust," "social," "celebrity," and other parts. A worthwhile future research point is to compare false and real messages with cognitively attractive features, or to evaluate how the features related to cognitive preferences contribute to the information virality.

In addition to studying the cognitive mechanisms at the data analysis level, we can also learn the mechanisms from the perspective of human brain cognitive functions. Advancements in neuroscience provide a promising way to study the cognitive mechanisms of false information. As stated by Poldrack et al. [130], the utilization of electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and other brain imaging tools can advance us in understanding how the human brain forms social behaviors. In addition, Adolphs [3] has identified the neural structures involved in the modulation of social cognition, such as the cingulate cortex, hippocampus, and basal forebrain. Arapakis et al. [6] utilize EEG recordings to measure user interests in news articles, and the experimental results show that the frontal alpha asymmetry (FFA) could objectively evaluate users' preference for media contents. To explain the mechanism of information virality, Scholz et al. [151] propose a fMRI data-based neurocognitive framework to evaluate users' willingness to share messages on Facebook. If we can understand the cognitive mechanisms of false information, then more effort can be focused on exploring the debunking information maximization methods to find robust counter measures for false information.

(2) ***Lack of standard datasets and benchmarks.*** Although researchers have done abundant works on FID, there is still a lack of benchmark datasets like *ImageNet* [31] for visual object recognition. Datasets, as a resource, are just as important as algorithms in FID. However, collecting fake messages is a time-consuming and labor-intensive process, which results in a lack of authoritative benchmarks.

We summarize the open datasets that have been used since 2015, as shown in Table 6, whose data is collected from *Sina Weibo* (e.g., RUMDECT[21] and Meida_Weibo[22]), *Twitter* (e.g., MediaEval,[23], PHEME,[24], RUMOUREVAL[25]), and other social platforms, as well as *snopes.com*, *politifact.com* (e.g., Emergent,[26] BuzzFeedWebis,[27] LIAR,[28] Declare,[29] and FakeNewsNet[30]) and other fact-checking sites. However, the annotation methods, data dimensions, and the ratio of true and false statements

[21] http://alt.qcri.org/~wgao/data/rumdect.zip.

[22] https://www.dropbox.com/s/xwlzvcxvws4m6ag/task3.zip?dl=0.

[23] https://github.com/MKLab-ITI/image-verification-corpus/tree/master/mediaeval2015.

[24] https://github.com/elkasrawi/Extended-Pheme-Dataset.

[25] https://figshare.com/articles/RumourEval_2019_data/8845580.

[26] https://github.com/willferreira/mscproject.

[27] https://github.com/BuzzFeedNews/2016-10-facebook-fact-check.

[28] https://www.cs.ucsb.edu/~william/data/liar_dataset.zip.

[29] https://www.mpi-inf.mpg.de/dl-cred-analysis/.

[30] https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset.

Table 6. A Summary of Widely Used Open Datasets

Dataset name	Data size		Information from datasets				Released year
	Fake	True	Content		Social context		
			Textual data	Visual data	User or web profiles	Interaction data	
MediaEval [12]	7,898 tweets	6,026 tweets	✓	✓	✓		2015
RUMDECT [106]	498 events	494 events	✓		✓	✓	2016
	2,313 events	2,351 events					
PHEME [205]	1,972 tweets	3,830 tweets	✓		✓	✓	2016
Emergent [42]	2,595 pieces of news related 300 events		✓			✓	2016
BuzzFeedWebis [134]	363 posts	1,264 posts	✓		✓		2016
LIAR [180]	12,836 pieces of news		✓		✓		2017
Media_Weibo [72]	Nearly 40k tweets with images		✓	✓	✓		2017
DeClare [133]	13,525 pieces of news		✓		✓	✓	2018
FakeNewsNet [161]	211 pieces of news	211 pieces of news	✓	✓	✓	✓	2018
RUMOUREVAL [53]	325 source tweets related to 9 events		✓		✓	✓	2019

among these datasets are different, which poses certain challenges for researchers to evaluate their model performance fairly. Shu et al. [162] have summarized the widely used evaluation metrics for FID, and the existing evaluation indicators are still *precision*, *recall*, *F1 score*, *accuracy*, and other machine learning model evaluation metrics. In FID, we need to define some more practical evaluation metrics. For example, in political elections, we would pay more attention to whether fake statements are more fully identified (i.e., more attention to *recall* than to *precision*), so using *F1 score* to evaluate the performance of detection models is not very appropriate. In future research, standard datasets and practical evaluation metrics are needed for comparing various FID algorithms and promoting the development of FID methods.

(3) ***Model adaptivity/generality to new events.*** FID methods should identify unseen, newly coming events, since the existing data of systems may differ from contents of emerging events. However, existing approaches tend to extract event-specific features that could hardly be shared with new events [204]. As stated by Tolosi et al. [173], feature engineering-based detection methods are hard to detect false information in different fields (such as politics, crimes, natural disasters), because features change dramatically across different events. Therefore, model generality or adaptivity is quite important to improve the robustness of FID models. Zubiaga et al. [206] state that the domain-dependent features' distribution could limit the generalization ability of models. As the distribution of most features directly correspond to events, the performance of FID models will be affected. Though we have discussed some crowd knowledge transfer models [138, 182, 187] in Section 4.2, there is much more to be investigated. Transfer learning models [59, 126], successfully used in other domains (such as sentiment classification [50], and image recognition [103]), can be leveraged to design domain-adaptative FID models. The usage of GAN-based discriminators [182] is another promising way to build generalized FID models with shared features.

Another interesting direction to be explored is that we should borrow knowledge from similar domains. For example, we can refer to web security, virus/spam detection methods [21, 153, 203], which also suffer from similar issues such as early detection and model generalization.

(4) ***Embracing of novel machine learning models.*** The FID process is by nature the learning of a classifier to identify the credibility of given claims. We have found that many studies build deep learning models [20, 72, 101, 106, 123, 144, 195] to improve the performance of automatic fact-checking. However, there are still more that can be explored. In the following, we present several representative examples that leverage advanced machine learning techniques to FID.

- ***Multi-task learning.*** Multi-task learning [109] is intended for improving the generalization performance of models by using domain knowledge contained in related tasks. Existing methods seek to find commonalities among multiple tasks by modeling the task relevance, such as feature sharing, sub-space sharing, and parameter sharing, as some supplementary knowledge for promoting learning effects of each task. For instance, Ma et al. [109] consider that the FID task is highly correlated with the stance classification task, so they propose a neural multi-task learning framework for better fact-checking. Under the mechanism of *weight sharing*, they present two RNN-based multi-task structures to jointly train the two tasks, which could extract ordinary as well as task-specific features for the rumor representation. Inspired by this work, we can investigate the connection and collaboration between FID and other tasks, and further design multi-task learning based algorithms to improve FID models performance.
- ***Few-shot learning.*** Few-shot learning [183] strives to addressing the data scarcity problem by leveraging few supervised information to recognize the samples from unseen classes. Existing few-shot learning methods usually decompose their training procedures into multiple meta-task learning procedures, similarly to meta learning [43], which extract transferable knowledge from different tasks' data. Consequently, this allows classification of new classes with only a small number of labeled data. To our best knowledge, there are less few-shot learning methods applied in FID, so we can learn from other related domains, such as text classification. To improve the induction and generalization of classifiers, Geng et al. [48] propose a dynamic routing algorithm-based classification architecture, called *Induction Networks*, which learns generalized class-level representations from a few samples. *Induction Networks* mainly contains an encoder module, an induction module, and a relation module. Specifically, the encoder module generates samples and the query representations, and then the induction module utilizes a transformation matrix to map the sample-level representations to class-level representations. Finally, the relation module calculates the matching degree between the query and each class. This work shows that few-shot learning has great potential in NLP, and we can continue studying few-shot learning-based FID methods.
- ***Semi-supervised models.*** Most existing FID works concentrate on supervised classification, and they usually train classifiers to identify false information through a large number of labeled data (e.g., fake or not). However, in many cases, we only have a small amount of labeled data. Semi-supervised models are often leveraged for dealing with the label sparsity issue. For example, Guacho et al. [55] propose a semi-supervised FID method, which leverages text embeddings based on tensor decomposition to capture the global and local features of social posts. After constructing the K-Nearest Neighbor (K-NN) graph of all the posts, they use a belief propagation algorithm to spread known labels into the graph for obtaining the final credibility of events. Furthermore, the development of graph neural networks also provides an opportunity for the research on semi-supervised detection models. GNNs, such as DeepWalk [129], LINE [171], and node2vec [54], utilize different sampling algorithms to generate the sequence of nodes, and then learns the representation of each node or propagation path by the skip-gram model. They introduce the *first-order proximity* (characterization of the similarity between two adjacent nodes) and *second-order proximity*

(characterization of the structural similarity between two nodes) into their loss functions to ensure that neural networks could fully extract features of graphs. Specially, GCNs [83], as discussed in Section 2, transfer information in adjacent convolutional layers through a non-linear transformation of the Laplacian matrix of graphs. Each convolutional layer only computes first-order proximity, so GCNs can learn high-level feature representations of nodes or propagation paths through multiple convolutional layers. In particular, GNNs are able to smooth label information through explicit graph regularization methods [184] for semi-supervised learning of graphs. Therefore, FID models could build information propagation graphs and combine GNNs to detect false information.

- ***Unsupervised models.*** If reliable unsupervised detection models can be directly established, then it is of great significance for the fast refutation of false information. Unsupervised models could evaluate the credibility of posts from the human–content interactions (such as publishing, or retweeting social media posts) and human–human interactions (such as following, or mentioning some users). On the one hand, the advancements of GAN, and VAE brings new possibilities for unsupervised FID models. However, PGMs can still play an important role in FID. For instance, Chen et al. [22] judge whether a post is fake or not from the users' posting behaviors. This unsupervised method utilizes AE to learn the latent representations of an individual's recent postings and their comments. When its reconstruction error converges, the model can be used to evaluate the credibility of new posts. If the model's reconstruction error exceeds a certain threshold, then this post may be a fake message. Yang et al. [191] consider the news truth and user credibility as latent variables, and utilize user reviews to infer their opinions on news authenticity. In other words, the truth of news depends on the credibility of users' opinions, and the credibility of opinions relies on the reputation of users. They utilize a Bayesian network to model the interaction process for inferring the truth of news articles without any labeled data. Actually, users' opinions may be influenced by other users, and their ability to identify false information on different topics is also different. These conditions could be further considered when using PGMs.
- ***Hybrid learning models.*** The development of hybrid learning models, combining linear models and deep learning models, has become a new research trend in AI, i.e., the combined usage of explicit features and latent features. It uses the complementarity of two types of learning models. For example, *Wide & Deep* [24] is a well-performed framework for recommender systems, where the *Wide* part extracts explicit features and the *Deep* part learns non-linear, latent features. There are also preliminary hybrid learning models in FID. Yang et al. [192] propose the *TI-CNN* model for detecting false information, which is trained with textual and visual information corporately based on the fusion of explicit and implicit feature spaces. Moreover, Zhang et al. [197] propose a FID model based on Bayesian deep learning, which uses LSTM to encode claims and user comments, and utilizes a Bayesian model to infer classification results. As hybrid learning models are still at its early stage, further research is needed in this direction, such as the fusion of probabilistic graph models and deep learning models.

(5) *Adversarial attack and defense in FID models.* Deep learning–based FID models contributes to an effective improvement of fact-checking performance. However, Szegedy et al. [169] have proven that the trained neural networks may fail to work against adversarial attacks, which means that adding some small perturbations to input vectors could make models get wrong results [4]. Existing FID studies rarely highlight the robustness of deep models that can be deceived by adversarial attacks.

Although few studies have been conducted on adversarial attack and defense in FID models, related works about other tasks (such as image classification [52, 169], speech recognition [15], text classification [86], and reinforcement learning [10]) has been investigated. Several works focus on the impact of adversarial attacks on models. For example, Dai et al. [28] present an adversarial attack method for graph data based on reinforcement learning (RL), which learns the optimal attack policy by increasing or decreasing the number of edges in graphs. To generate universal adversarial perturbations for text, Behjati et al. [9] propose a gradient projection-based attack method. Jia et al. [71] attack Q&A systems by adding sentences or phrases to questions that do not cause difficulty to human understanding.

Above attack studies could guide the research of adversarial attack defense in FID models. Zhou et al. [202] further divide adversarial attacks on FID models into *fact distortion*, *subject-object exchange*, and *cause confounding*. To resist adversarial attacks, they further propose a crowdsourced knowledge graph to collect timely facts about news events. Qiu et al. [139] classify defensive methods into three categories, including *modifying data* (e.g., adversarial training and gradient hiding), *modifying models* (e.g., regularization and defensive distillation), and *using auxiliary tools* (e.g., *Defense-GAN* [148]). Whether it is the attack on models or the manipulation of data, higher requirements are put forward for the robustness of FID systems. Accordingly, there are still more efforts to be conducted on adversarial attack and defense on FID.

(6) *Explanatory detection models.* Providing evidence or explanations of decision results can increase user trust in detection models. Though there have been few works on explanatory FID models, the application of explanations has been investigated in other related domains, such as recommender systems.

Explainable recommendation, which provides explanations about why an item is recommended, has attracted increasing attention in recent years [198]. It can improve users' acceptability, credibility, and satisfaction with recommender systems and enhance the systems' persuasiveness. For example, Chen et al. [23] present a visually explainable recommendation method based on attentive neural networks to model user attention on images. Users can understand why a product is recommended by providing personalized and intuitive visual highlights. Catherine et al. [18] study how to generate explainable recommendations with the support of external knowledge graphs, and they propose a personalized PageRank procedure to rank items and knowledge graph entities together. The work of Wang et al. [181] proposes a model-agnostic explanatory recommendation system based on reinforcement learning (RL), which can flexibly control the presentation quality of explanations. Above all, the methods used in such explainable recommendation systems can inspire us to design better explainable FID systems.

From a higher perspective, machine learning models have powered breakthroughs in diverse application areas (beyond recommender systems and FID). Despite the big success, we still lack understanding of their intrinsic behaviors, such as how a classifier arrives at a particular decision. This has resulted in the surging research direction of Interpretable Machine Learning (IML). IML gives machine learning models the ability to explain or present in human understandable terms [2, 34]. Du et al. [35] define two types of interpretability: *model-level interpretation* and *prediction-level explanation*. Model-level interpretation, for increasing the transparency of models themselves, can illuminate the inner working mechanisms of machine learning models. Prediction-level explanation helps uncover the relations between specific inputs and model outputs. For FID, it pays more attention to prediction-level explanation, which can illustrate how a decision can be arrived (using elements such as source reliability, evidence, and stance). A representative scheme of constructing prediction-level explainable models is employing attention mechanisms, which is widely utilized to explain decision results made by sequential models (e.g., RNNs). We should also study other approaches rooting from IML to enhance the explainability of FID systems.

(7) *Aggregation of crowd wisdom.* How to aggregate crowd wisdom is important for FID systems, because crowd contribution data often has noise. Most users' opinions can be effectively used for identifying false information, but there is also a situation where the truth is in the hands of the minority. Therefore, it is still necessary to explore the aggregation and optimization method of crowd wisdom for FID in the future.

We can learn from truth discovery systems. With the ability to extract reliable information from conflicting multi-sourced data using human intelligence, truth discovery has become an increasingly significant research topic. For FID, we also have multiple posts about an event, and the target is to identify the truth of this event. Therefore, there are similarities between the two research problems, and we can borrow knowledge from truth discovery systems to facilitate the FID. For example, Liu et al. [98] propose an expert validation-assisted image label truth discovery method, aiming at deducing correct labels as much as possible from noisy crowdsourced labels. In particular, it utilizes a semi-supervised learning algorithm in the manner of human–machine collaboration that can maximize the influence of expert labels and reduce efforts of experts. Zhang et al. [196] propose a probabilistic graph-based truth discovery model named "*TextTruth*," which selects highly trustworthy answers to questions by comprehensively learning the trustworthiness of key factors (a group of keywords). The *TextTruth* infers the credibility of answer providers and the trustworthiness of answer factors together in an unsupervised manner. Yin et al. [193] present a model of crowd wisdoms aggregation in an unsupervised manner, called Label-Aware Autoencoders (*LAA*), which extracts underlying features and patterns of multi-sourced labels and infers the trustworthy labels by a classifier and a reconstructor. To tackle the challenge that the same information source has different credibility on various topics, Ma et al. [105] propose a crowdsourced data aggregation method named *FaitCrowd*. The *FaitCrowd* jointly learns the topic distribution of questions, topic-based knowledge of answer providers, and true answers by modeling question contents and answers from publishers together on a probabilistic Bayesian model.

(8) *Propagation by social bots.* Existing FID studies concentrate on the contents and propagation patterns of claims. The characters of the "accounts" that publish and disseminate posts, however, are not well investigated. Recently, several efforts have been made to study the root causes of false information spreading as rapidly as viruses. For example, Shao et al. [155] perform a detailed analysis of 14 million tweets during the 2016 U.S. presidential election, and they observe that the "social bots" apparently facilitate the rapid diffusion of false information. A social bot usually refers to a computer algorithm or software program that imitates human interaction behaviors (e.g., producing contents, following other accounts, retweeting posts, etc.) for some purpose [40]. These malicious bot accounts are abnormally active in very early stages of fake tweets dissemination. Besides, after modeling social interactions and emotional interactions of social bots, Stella et al. [167] find that they increase the exposure of negative and violent contents on social networks.

Above findings suggest that the suppression of social bots can be a promising way to mitigate the dissemination of false information. Some researchers have analyzed behavior patterns of social bots and proposed some detection methods. For example, Ferrara et al. [40] classify existing social bots detection approaches into four categories, includiing graph-based models, crowdsourcing, feature-based models, and hybrid models. Almaatoug et al. [5] design a social bots detection method that incorporates content attributes, social interactions, and profile properties. Similarly, Minnich et al. [115] propose the *BotWalk* detection method, which utilizes several features to distinguish users from bot accounts, such as metadata, content, temporal information, and network interaction. Cresci et al. [27] conduct a penetrating analysis of collective behaviors of social bots and introduce a *Social Fingerprinting* technique for spambot detection. In particular, they exploit the *digital DNA* technique to characterize collective behaviors of all the accounts, and then they propose a DNA-inspired method to identify genuine accounts and spambots. Cresci et al. [26] also leverage characteristics of group accounts to detect malicious bots. As social bots promote the spread of low-credibility statements and the exposure of negative content [155, 167], future works could combine FID with social bot detection, providing new solutions for the fast refutation of false information.

(9) ***False Information Mitigation.*** Effective FID is a part of the prevention of false information, and it also needs scientific research to reduce the impact of false information, which belongs to the research scope of false information mitigation. Some works have reviewed the approaches of false information mitigation and intervention. For example, Sharma et al. [156] summarize three kinds of mitigation methods from the perspective of information diffusion, namely "*decontamination,*" "*competition cascades,*" and "*multi-stage interference.*" Shu et al. [159] divide existing mitigation strategies into "*user identification,*" "*network size estimation,*" and "*network intervention.*" As each user plays a different role in the dissemination of false information, such as opinion leaders, guardians, malicious disseminators and onlookers, it is necessary to take flexible mitigation measures. For instance, opinion leaders and guardians are suitable to be recommended with factual information to help spread the truth [175], while malicious accounts or bots should be curbed [122]. As Ozturk et al. [125] once stated, displaying fake messages with fact-checking information on Twitter contributes to the reduction of continual dissemination of false information. Based on this observation, Budak et al. [13] propose the Multi-Campaign Independence Cascade Model, which contains a campaign of false information and a campaign of true information. Furthermore, we can also utilize the Multivariate Hawkes Process [37] to model the propagation dynamics of false information under the influence of external interventions.

In future research, FID can be combined with above mitigation strategies to explore more promising works in preventing the dissemination of false information on social networks. Moreover, Sundar [168] once confirmed that the existence of source attribution in social posts improved users' perception of the credibility and quality of online information. Consequently, source attribution and causal inference [158] can also be used to guide the detection of false information on social media.

6 CONCLUSION

We have made a systematic review of the research trends of FID. Having given a brief review of the literature of FID, we present several new research challenges and techniques of it, including early detection, detection by multimodal data fusion, and explanatory detection. We further investigate the usage of crowd intelligence in FID, including crowd intelligence–based FID models and hybrid human–machine FID models. Though there has been a big research progress in FID, it is still at the early stage and there are numerous open research issues and promising research directions to be studied, such as model adaptivity/generality to new events, embracing novel machine learning models, explanatory detection models, and so on.

REFERENCES

[1] Alberto Acerbi. 2019. Cognitive attraction and online misinformation. *Palgrave Commun.* 5, 1 (2019), 15.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] R. Adlolphs. 2003. Cognitive neuroscience of human social behavior. *Nat. Rev. Neurosci.* 4, 3 (2003), 165–178.

[4] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6 (2018), 14410–14430.

[5] Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K. Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. 2016. If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts. *Int. J. Inf. Secur.* 15, 5 (2016), 475–491.

[6] Ioannis Arapakis, Miguel Barreda-Angeles, and Alexandre Pereda-Baños. 2017. Interest as a proxy of engagement in news reading: Spectral and entropy analyses of EEG activity patterns. *IEEE Trans. Affect. Comput.* 10, 1 (2017), 100–114.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Arxiv Preprint Arxiv:1409.0473* (2014).

[8] Jonathan Bakdash, Char Sample, Monica Rankin, Murat Kantarcioglu, Jennifer Holmes, Sue Kase, Erin Zaroukian, and Boleslaw Szymanski. 2018. The future of deception: Machine-generated and manipulated images, video, and audio?. In *Proceedings of the 2018 International Workshop on Social Sensing (SocialSens)*. IEEE, 2–2.

[9] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7345–7349.

[10] Vahid Behzadan and Arslan Munir. 2017. Vulnerability of deep reinforcement learning to policy induction attacks. In *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 262–275.

[11] Sreyasee Das Bhattacharjee, Ashit Talukder, and Bala Venkatram Balantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 556–565.

[12] Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. 2014. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 743–748.

[13] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. 2011. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*. 665–674.

[14] Juan Cao, Junbo Guo, Xirong Li, Zhiwei Jin, Han Guo, and Jintao Li. 2018. Automatic rumor detection on microblogs: A survey. *Arxiv Preprint Arxiv:1807.03505* (2018).

[15] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security'16)*. 513–530.

[16] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 675–684.

[17] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Res.* 23, 5 (2013), 560–588.

[18] Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi, and William Cohen. 2017. Explainable entity-based recommendations with knowledge graphs. *Arxiv Preprint Arxiv:1707.05254* (2017).

[19] Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An attentive survey of attention models. *Arxiv Preprint Arxiv:1904.02874* (2019).

[20] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 40–52.

[21] Wenji Chen, Yang Liu, and Yong Guan. 2013. Cardinality change-based early detection of large-scale cyber-attacks. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'13)*. IEEE, 1788–1796.

[22] Weiling Chen, Yan Zhang, Chai Kiat Yeo, Chiew Tong Lau, and Bu Sung Lee. 2018. Unsupervised rumor detection based on users' behaviors using neural networks. *Pattern Recogn. Lett.* 105 (2018), 226–233.

[23] Xu Chen, Yongfeng Zhang, Hongteng Xu, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Visually explainable recommendation. *Arxiv Preprint Arxiv:1801.10288* (2018).

[24] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. ACM, 7–10.

[25] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Arxiv Preprint Arxiv:1406.1078* (2014).

[26] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 963–972.

[27] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Trans. Depend. Sec. Comput.* 15, 4 (2017), 561–576.

[28] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial attack on graph structured data. *Arxiv Preprint Arxiv:1806.02371* (2018).

[29] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*. 3844–3852.

[30] Marco L. Della Vedova, Eugenio Tacchini, Stefano Moret, Gabriele Ballarin, Massimo DiPierro, and Luca de Alfaro. 2018. Automatic online fake news detection combining content and social signals. In *Proceedings of the 2018 22nd Conference of Open Innovations Association (FRUCT'18)*. IEEE, 272–279.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[32] Manqing Dong, Lina Yao, Xianzhi Wang, Boualem Benatallah, Quan Z. Sheng, and Hao Huang. 2018. DUAL: A deep unified attention model with latent relation representations for fake news detection. In *International Conference on Web Information Systems Engineering*. Springer, 199–209.

[33] Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and Guohui Li. 2019. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 569–578.

[34] Finale Doshi-Velez and Been Kim. 2018. Considerations for evaluation and generalization in interpretable machine learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 3–17.

[35] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[36] Don Fallis. 2015. What is disinformation? *Library Trends* 63, 3 (2015), 401–426.

[37] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. Fake news mitigation via point process based intervention. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 1097–1106.

[38] Alireza Farasat, Alexander Nikolaev, Sargur N. Srihari, and Rachael Hageman Blair. 2015. Probabilistic graphical models in modern social network analysis. *Soc. Netw. Anal. Min.* 5, 1 (2015), 62.

[39] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 595–602.

[40] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[41] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. 2012. Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Trans. Inf. Forens. Secur.* 7, 5 (2012), 1566–1577.

[42] William Ferreira and Andreas Vlachos. 2016. Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1163–1168.

[43] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 1126–1135.

[44] Luciano Floridi. 2018. Artificial intelligence, deepfakes and a future of ectypes. *Philos. Technol.* 31, 3 (2018), 317–321.

[45] Jessica Fridrich and Jan Kodovsky. 2012. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forens. Secur.* 7, 3 (2012), 868–882.

[46] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

[47] Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. ExFaKT: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, 87–95.

[48] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3895–3904.

[49] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th international conference on artificial intelligence and statistics*. 249–256.

[50] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 513–520.

[51] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[52] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *Arxiv Preprint Arxiv:1412.6572* (2014).

[53] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 845–854.

[54] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[55] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'18)*. IEEE, 322–325.

[56] David Güera and Edward J. Delp. 2018. Deepfake video detection using recurrent neural networks. In *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'18)*. IEEE, 1–6.

[57] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Sci. Adv.* 5, 1 (2019), eaau4586.

[58] Bin Guo, Chao Chen, Daqing Zhang, Zhiwen Yu, and Alvin Chin. 2016. Mobile crowd sensing and computing: When participatory sensing meets participatory social media. *IEEE Commun. Mag.* 54, 2 (2016), 131–137.

[59] Bin Guo, Jing Li, Vincent W. Zheng, Zhu Wang, and Zhiwen Yu. 2018. Citytransfer: Transferring inter-and intra-city knowledge for chain store site recommendation based on multi-source urban data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (2018), 135.

[60] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 943–951.

[61] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 729–736.

[62] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, 153–164.

[63] Andreas Hanselowski, Avinesh P. V. S., Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *Arxiv Preprint Arxiv:1806.05180* (2018).

[64] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.

[65] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[66] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. 2016. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 495–503.

[67] Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.

[68] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2014. Social spammer detection with sentiment information. In *Proceedings of the 2014 IEEE International Conference on Data Mining*. IEEE, 180–189.

[69] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 101–117.

[70] Jun Ito, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. 2015. Assessment of tweet credibility with LDA features. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 953–958.

[71] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *Arxiv Preprint Arxiv:1707.07328* (2017).

[72] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 795–816.

[73] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International Conference on Social Computing, Behavioral-cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 14–24.

[74] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *Proceedings of the 2014 IEEE International Conference on Data Mining*. IEEE, 230–239.

[75] Zhiwei Jin, Juan Cao, Jiebo Luo, and Yongdong Zhang. 2016. Image credibility analysis with effective domain transferred deep networks. *Arxiv Preprint Arxiv:1611.05328* (2016).

[76] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

[77] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multimedia* 19, 3 (2016), 598–608.

[78] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the World Wide Web Conference*. ACM, 2915–2921.

[79] Jooyeon Kim, Dongkwan Kim, and Alice Oh. 2019. Homogeneity-based transmissive process to model true and false news in social networks. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining*. ACM, 348–356.

[80] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 324–332.

[81] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Arxiv Preprint Arxiv:1408.5882* (2014).

[82] Diederik P. Kingma and Max Welling. 2014. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations (ICLR'14)*, Vol. 19.

[83] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *Arxiv Preprint Arxiv:1609.02907* (2016).

[84] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[85] Pavel Korshunov and Sébastien Marcel. 2019. Vulnerability assessment and detection of deepfake videos. In *2019 International Conference on Biometrics (ICB'19)*. IEEE, 1–6.

[86] Volodymyr Kuleshov, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. 2018. Adversarial examples for natural language classification problems. https://openreview.net/forum.

[87] K. P. Krishna Kumar and G. Geethakumari. 2014. Detecting misinformation in online social networks using cognitive psychology. *Human-centric Comput. Inf. Sci.* 4, 1 (2014), 14.

[88] Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V. S. Subrahmanian. 2018. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 333–341.

[89] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *Arxiv Preprint Arxiv:1804.08559* (2018).

[90] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*. IEEE, 1103–1108.

[91] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[92] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *J. Appl. Res. Mem. Cogn.* 6, 4 (2017), 353–369.

[93] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Publ. Interest* 13, 3 (2012), 106–131.

[94] Lizhao Li, Guoyong Cai, and Nannan Chen. 2018. A rumor events detection method based on deep bidirectional GRU neural network. In *Proceedings of the 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC'18)*. IEEE, 755–759.

[95] Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1173–1179.

[96] Wei Li, Wen-jun Wu, Huai-min Wang, Xue-qi Cheng, Hua-jun Chen, Zhi-hua Zhou, and Rong Ding. 2017. Crowd intelligence in AI 2.0 era. *Front. Inf. Technol. Electr. Eng.* 18, 1 (2017), 15–43.

[97] Wee Yong Lim, Mong Li Lee, and Wynne Hsu. 2017. iFACT: An interactive framework to assess claims from tweets. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 787–796.

[98] Mengchen Liu, Liu Jiang, Junlin Liu, Xiting Wang, Jun Zhu, and Shixia Liu. 2017. Improving learning-from-crowds through expert validation. In *Proceedings of the International Joint Conferences on Artficial Intelligence (IJCAI'17)*. 2329–2336.

[99] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. 2018. Mining significant microblogs for misinformation identification: An attention-based approach. *ACM Trans. Intell. Syst. Technol.* 9, 5 (2018), 50.

[100] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1867–1870.

[101] Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedingsof the 32nd AAAI Conference on Artificial Intelligence*.

[102] Yang Liu and Songhua Xu. 2016. Detecting rumors through modeling information propagation networks in a social media environment. *IEEE Trans. Comput. Soc. Syst.* 3, 2 (2016), 46–62.

[103] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*. 97–105.

[104] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 252–256.

[105] Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 745–754.

[106] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the International Conferences on Artificial Intelligence (IJCAI'16)*. 3818–3824.

[107] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1751–1754.

[108] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 708–717.

[109] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion of The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 585–593.

[110] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1980–1989.

[111] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *Proceedings of the World Wide Web Conference*. ACM, 3049–3055.

[112] Scott McCloskey and Michael Albright. 2019. Detecting GAN-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP'19)*. IEEE, 4584–4588.

[113] Pietro Michelucci and Janis L. Dickinson. 2016. The power of crowds. *Science* 351, 6268 (2016), 32–33.

[114] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

[115] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of twitter bot networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 467–474.

[116] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.* 17, 3 (2017), 26.

[117] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *Arxiv Preprint Arxiv:1902.06673* (2019).

[118] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 441–450.

[119] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B. S. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. 2019. Detecting GAN generated fake images using co-occurrence matrices. *Electronic Imaging* 2019, 5 (2019), 532-1–532-7.

[120] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 189–199.

[121] An T. Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[122] Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. 213–222.

[123] Tu Ngoc Nguyen, Cheng Li, and Claudia Niederée. 2017. On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners. In *International Conference on Social Informatics*. Springer, 141–158.

[124] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing*.

[125] Pinar Ozturk, Huaye Li, and Yasuaki Sakamoto. 2015. Combating rumor spread on social media: The effectiveness of refutation and warning. In *Proceedings of the 2015 48th Hawaii International Conference on System Sciences.* IEEE, 2406–2414.

[126] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2009), 1345–1359.

[127] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web* 8, 3 (2017), 489–508.

[128] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *Arxiv Preprint Arxiv:1708.07104* (2017).

[129] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 701–710.

[130] Russell A. Poldrack and Martha J. Farah. 2015. Progress and challenges in probing the human brain. *Nature* 526, 7573 (2015), 371.

[131] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, 1003–1012.

[132] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Companion of the The Web Conference 2018 on The Web Conference 2018.* International World Wide Web Conferences Steering Committee, 155–158.

[133] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. *Arxiv Preprint Arxiv:1809.06416* (2018).

[134] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *Arxiv Preprint Arxiv:1702.05638* (2017).

[135] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.* 51, 5 (2018), 92.

[136] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 1589–1599.

[137] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM'19).* IEEE, 518–527.

[138] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural user response generator: Fake news detection with collective user intelligence. In *Proceedings of the Internaional Conferences on Artificial Intelligence (IJCAI'18).* 3834–3840.

[139] Shilin Qiu, Qihe Liu, Shijie Zhou, and Chunjiang Wu. 2019. Review of artificial intelligence adversarial attack and defense technologies. *Appl. Sci.* 9, 5 (2019), 909.

[140] Xiaoyan Qiu, Diego F. M. Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. 2017. Limited individual attention and online virality of low-quality information. *Nat. Hum. Behav.* 1, 7 (2017), 0132.

[141] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2931–2937.

[142] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 985–994.

[143] Victoria L. Rubin and Tatiana Lukoianova. 2015. Truth and deception at the rhetorical structure level. *J. Assoc. Inf. Sci. Technol.* 66, 5 (2015), 905–917.

[144] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* ACM, 797–806.

[145] Derek Ruths. 2019. The misinformation machine. *Science* 363, 6425 (2019), 348–348.

[146] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference.* ACM, 1337–1345.

[147] Ronald Salloum, Yuzhuo Ren, and C.-C. Jay Kuo. 2018. Image splicing localization using a multi-task fully convolutional network (MFCN). *J. Vis. Commun. Image Represent.* 51 (2018), 201–209.

[148] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605* (2018).

[149] Justin Sampson, Fred Morstatter, Liang Wu, and Huan Liu. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* ACM, 2377–2382.

[150] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proc. Natl. Acad. Sci. U.S.A.* 116, 16 (2019), 7662–7669.

[151] Christin Scholz, Elisa C. Baek, Matthew Brook O'Donnell, Hyun Suk Kim, Joseph N. Cappella, and Emily B. Falk. 2017. A neural model of valuation and information virality. *Proc. Natl. Acad. Sci. U.S.A.* 114, 11 (2017), 2881–2886.

[152] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 815–823.

[153] Suranga Seneviratne, Aruna Seneviratne, Mohamed Ali Kaafar, Anirban Mahanti, and Prasant Mohapatra. 2015. Early detection of spam mobile apps. In *Proceedings of the 24th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 949–959.

[154] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web.* 745–750.

[155] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nat. Commun.* 9, 1 (2018), 4787.

[156] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3 (2019), 1–42.

[157] Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web.* International World Wide Web Conferences Steering Committee, 101–102.

[158] Richard M. Shiffrin. 2016. Drawing causal inference from big data. *Proc. Natl. Acad. Sci. U.S.A.* 113, 27 (2016), 7308–7309.

[159] Kai Shu, H. Russell Bernard, and Huan Liu. 2019. Studying fake news via network analysis: Detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining.* Springer, 43–65.

[160] Kai Shu, Deepak Mahudeswaran, and Huan Liu. 2019. FakeNewsTracker: A tool for fake news collection, detection, and visualization. *Comput. Math. Organiz. Theory* 25, 1 (2019), 60–71.

[161] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media. *Big Data* 8, 3 (2020), 171–188.

[162] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newslett.* 19, 1 (2017), 22–36.

[163] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting tri-relationship for fake news detection. *Arxiv Preprint Arxiv:1712.07709* (2017).

[164] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'18).* IEEE, 430–435.

[165] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining.* ACM, 312–320.

[166] Bernd Carsten Stahl. 2006. On the difference or equality of information, misinformation, and disinformation: A critical research perspective. *Inf. Sci.* 9 (2006), 83–96.

[167] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440.

[168] S. Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journal. Mass Commun. Quart.* 75, 1 (1998), 55–68.

[169] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *Arxiv Preprint Arxiv:1312.6199* (2013).

[170] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *Arxiv Preprint Arxiv:1704.07506* (2017).

[171] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 1067–1077.

[172] Jeffrey E. Thomas. 1986. Statements of fact, statements of opinion, and the first amendment. *Calif. L. Rev.* 74 (1986), 1001.

[173] Laura Tolosi, Andrey Tagarev, and Georgi Georgiev. 2016. An analysis of event-agnostic features for rumour classification in twitter. In *Proceedings of the 10th International AAAI Conference on Web and Social Media.*

[174] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 517–524.

[175] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 275–284.

[176] Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 575–583.

[177] Luis Von Ahn. 2008. Human computation. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. IEEE Computer Society, 1–2.

[178] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 321, 5895 (2008), 1465–1468.

[179] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[180] William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Arxiv Preprint Arxiv:1705.00648* (2017).

[181] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018. A reinforcement learning framework for explainable recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 587–596.

[182] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.

[183] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* 53, 3 (2020), 63.

[184] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. 2012. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*. Springer, 639–655.

[185] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 6004 (2010), 686–688.

[186] Ke Wu, Song Yang, and Kenny Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *Proceedings of the 2015 IEEE 31st International Conference on Data Engineering*. IEEE, 651–662.

[187] Liang Wu, Jundong Li, Xia Hu, and Huan Liu. 2017. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 99–107.

[188] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 637–645.

[189] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 823–831.

[190] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia Ben Hu. 2019. XFake: Explainable fake news detector with visualizations. In *Proceedings of the World Wide Web Conference*. ACM, 3600–3604.

[191] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence*.

[192] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S. Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *Arxiv Preprint Arxiv:1806.00749* (2018).

[193] Li'ang Yin, Jianhua Han, Weinan Zhang, and Yong Yu. 2017. Aggregating crowd wisdoms with label-aware autoencoders. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 1325–1331.

[194] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Comput. Secur.* 83 (2019), 106–121.

[195] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Attention-based convolutional approach for misinformation identification from massive and noisy microblog posts. *Computer and Security* 83 (2019), 106–121.

[196] Hengtong Zhang, Yaliang Li, Fenglong Ma, Jing Gao, and Lu Su. 2018. TextTruth: An unsupervised approach to discover trustworthy information from multi-sourced text data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2729–2737.

[197] Qiang Zhang, Aldo Lipani, Shangsong Liang, and Emine Yilmaz. 2019. Reply-aided detection of misinformation via bayesian deep learning. In *Proceedings of the World Wide Web Conference*. ACM, 2333–2343.

[198] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *Arxiv Preprint Arxiv:1804.11192* (2018).

[199] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1395–1405.

[200] Shi Zhi, Yicheng Sun, Jiayi Liu, Chao Zhang, and Jiawei Han. 2017. ClaimVerif: A real-time claim verification system using the web and fact databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2555–2558.

[201] Xinyi Zhou and Reza Zafarani. 2018. Fake News: A survey of research, detection methods, and opportunities. *Arxiv Preprint Arxiv:1812.00315* (2018).

[202] Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via NLP is vulnerable to adversarial attacks. *Arxiv Preprint Arxiv:1901.09657* (2019).

[203] Cliff C. Zou, Weibo Gong, Don Towsley, and Lixin Gao. 2005. The monitoring and early detection of internet worms. *IEEE/ACM Trans. Netw.* 13, 5 (2005), 961–974.

[204] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.* 51, 2 (2018), 32.

[205] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *Arxiv Preprint Arxiv:1610.07363* (2016).

[206] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*. Springer, 109–123.

Received August 2019; revised March 2020; accepted April 2020