Fundamental Research 5 (2025) 332–346

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Review

An overview of fake news detection: From a new perspective

Bo Hu, Zhendong Mao*, Yongdong Zhang*

School of Information Science and Technology, University of Science and Technology of China, Hefei 230022, China

ARTICLE INFO

Article history:
Received 28 January 2023
Received in revised form 17 October 2023
Accepted 21 January 2024
Available online 22 February 2024

Keywords:
Fake news detection
Social media
Intentional creation
Heteromorphic transmission
Controversial reception

ABSTRACT

With the rapid development and popularization of Internet technology, the propagation and diffusion of information become much easier and faster. While making life more convenient, the Internet also promotes the wide spread of fake news, which will have a great negative impact on countries, societies, and individuals. Therefore, a lot of research efforts have been made to combat fake news. Fake news detection is typically a classification problem aiming at verifying the veracity of news contents, which may include texts, images and videos. This article provides a comprehensive survey of fake news detection. We first summarize three intrinsic characteristics of fake news by analyzing its entire diffusion process, namely intentional creation, heteromorphic transmission, and controversial reception. The first refers to why users publish fake news, the second denotes how fake news propagates and distributes, and the last means what viewpoints different users may hold for fake news. We then discuss existing fake news detection approaches according to these characteristics. Thus, this review will enable readers to better understand this field from a new perspective. We finally discuss the trends of technological advances in this field and also outline some potential directions for future research.

1. Introduction

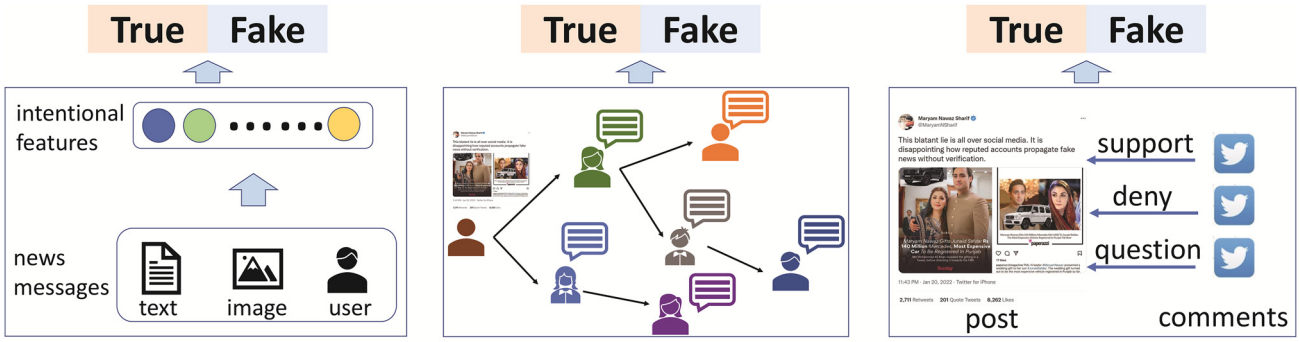
With the development of Internet communication technology and the rise of social networks, it becomes possible for ordinary people to publish news and make comments online, which even though brings great convenience, while provides an environment conducive to the creation and spread of fake news. Fake news may have a negative impact on countries, societies, and individuals. As for the countries, a large amount of fake news arose during the U.S. 2016 presidential election [1], which might have heavily influenced the election results. As for societies, fake news sometimes emerges with natural disasters and pandemics, such as the Japan earthquake in 2011 [2], Hurricane Sandy in 2012 [3], and the COVID-19 pandemic in 2019 [4], which may cause panics among the public. As for individuals, fake news which claimed that Obama was injured in the blast has resulted in a dramatic collapse in the stock market [5], which might damage the property of individuals. In addition, a lot of health misinformation regarding COVID-19 and vaccination was disseminated during the pandemic, which can even harm the physical well-being of deceived individuals.

In the past few years, there have been a lot of attempts to distinguish fake news from real news. Famous social networks like Twitter, Facebook and Weibo have developed anti-rumor centers, which allow users to report and dispel possible fake news. Such a mechanism, to some extent, reduces the negative impact of fake news, but is inefficient as it relies on extensive manual review and expert knowledge. More importantly, it is unable to detect emerging fake news at an early stage, and thus fails to minimize the damage caused by fake news. To address the above issues, various automatic fake news detection algorithms have been developed which can detect fake news as early as possible and help stop the viral spread of such news. Early studies mainly focus on designing hand-crafted features, e.g., statistical features [3,6–11], topic features [12,13], lexical features [10,13–16], and syntactic features [17–19], and then training supervised [6,7,18,20–23] or unsupervised [24,25] classifiers to distinguish between fake and real news. Recent studies investigate the effectiveness of propagation patterns on fake news detection, and a variety of propagation tree or propagation graph based models [3,12,14,26] have been proposed and applied successively. More recently with the development of deep learning [27], a growing number of studies [19,20,28–44] are moving towards exploiting deep neural networks to extract features or model propagation patterns.

There have been previous reviews of fake news detection techniques. For instance, Zubiaga et al. [45] provide an overview of existing research with the goal of developing a classification system of fake news, including four components, such as detection, tracking, stance classification and veracity classification. Zhou et al. [5] categorize current techniques into knowledge-based methods, style-based approaches, propagation-based algorithms and credibility-based networks. Zhang et al. [46] review the fake news detection mechanisms corresponding to three types of features of fake news, such as creator/user-based, news

* Corresponding authors.
E-mail addresses: zdmao@ustc.edu.cn (Z. Mao), zhyd73@ustc.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.fmre.2024.01.017



a Intentional feature based approach b Propagation based approach c Stance based approach

Fig. 1. Three categories of fake news detection approaches based on three characteristics: intentional creation, heteromorphic transmission, and controversial reception. (a) Intentional feature-based approaches first extract features to describe intentions of news messages, and then use these features for classification. (b) Propagation-based approaches first construct the propagation structures, and then study the structure patterns and information diffusion for veracity evaluation. (c) Stance-based approaches exploit the stances of different users as clues to facilitate fake news detection.

content-based and social context-based features. Shu et al. [47] discuss fake news detection from a data mining perspective, including feature extraction and model construction. Varshney et al. [48] review the prior works on the procedure of fake news detection, including social media data collection, preprocessing, feature analysis and detection models. Rohera et al. [49] classify previous works into different categories, such as supervised learning, semi-supervised learning and unsupervised learning, etc. Schlicht et al. [50] propose a taxonomy with 6 dimensions, such as inputs, source, topics, types, tasks, and detection methods, to categorize and discuss recent research works on health misinformation detection. Chen et al. [51] provide a systematic review on health misinformation detection from the perspectives of misinformation characterization, detection mechanisms, and intervention efforts made by individuals, organizations and governments.

In this paper, we provide a thorough survey of fake news detection from a brand-new perspective: the intrinsic characteristics of fake news. Specifically, by analyzing the entire diffusion process of fake news, we summarize its three main characteristics, i.e., intentional creation, heteromorphic transmission, and controversial reception, detailed as follows.

(1) **Intentional creation**. Unlike real news that reports real events, fake news is usually created with intention, e.g., to mislead the public or manipulate opinions [47]. For example, an art painting from 2009 was posted to mislead the public, which indicated that Ronald McDonald was flooded in Hurricane Sandy [52].

(2) **Heteromorphic transmission**. Real news is usually spread by ordinary users, while fake news however tends to be propagated by diverse types of users, which thereby results in heteromorphic transmission patterns. For example, as pointed out by Ma et al. [26], fake news is typically posted by a low-impact user first and then widely spread by opinion leaders, while real news is usually initiated by an opinion leader and spread by normal users.

(3) **Controversial reception**. People are more likely to hold different views towards fake news than real news. Given a piece of fake news, some people believe it, while others may suspect or deny its veracity. For example, users can express their attitudes through “thumbs up” or “thumbs down” on Facebook [47].

Based on these characteristics, we categorize existing techniques into three groups: intentional feature-based, propagation-based, and stance-based approaches, as shown in Fig. 1.

- *Intentional Feature-based* approaches first extract features to describe intentions of news messages, and then use these features for classification.
- *Propagation-based* approaches first construct the propagation structures, and then study the structure patterns and information diffusion for veracity evaluate.
- *Stance-based* approaches exploit the stances of different users as clues to facilitate fake news detection.

Thus, this survey can enable readers to understand this field from a new perspective, and help us reveal the trends of technological advances, which provides insights on how to design effective and explainable detection mechanisms, including

- **Characteristic Selection:** Characteristics from all three categories of the diffusion process tend to be utilized together for fake news detection.
- **Framework Design:** A framework can be designed to capture the characteristics from all three categories for effective fake news detection.
- **Result Explanation:** Detection results can be explained in a more fine-grained manner, which reveals the key factors of fake news.

The rest of this survey is organized as follows. We first introduce the definition of fake news and briefly describe the fake news detection task in Section 2. Then we discuss three intrinsic characteristics of fake news and corresponding detection approaches in Sections 3, 4, 5, respectively. After that, we present popular evaluation datasets in Section 6, and discussion and future directions in Section 7. Finally, we conclude in section Section 8.

2. Definitions

Fake News: There is no widely accepted definition of fake news. In this survey, we follow the narrow definition of fake news used in [47,53]: fake news is intentionally created and verified false. The key distinctive features of fake news are authenticity and intentionality. It is easy to differentiate fake news from other related concepts by these two features. For example, when the authenticity is unverified and the intention is unknown, the concept denotes rumor [54]; and when the authenticity is false, but the intention is not bad, the concept denotes misinformation [5,47]. Note that fake news is a special case of disinformation [5]. Fake news is usually limited to news articles while disinformation includes all kinds of information. In this survey, our main focus is on the work of fake news detection. However, the literature indicates that the detection of rumors and misinformation exhibits analogous characteristics to fake news, and the detection methods can also be applied to fake news detection. Therefore, this paper encompasses representative methods in rumors and misinformation detection for more comprehensive overview of fake news detection.

Fake News Detection: Given an event $x \in \mathcal{X}$, where $\mathcal{X}$ is the event set, and a set of event-related news messages $\mathcal{M}_x = \{m_1, m_2, \ldots, m_N\}$. Each message $m_i \in \mathcal{M}_x$ contains textual descriptions and visual contents, regarding event x. Let $\mathcal{U} = \{u_1, u_2, \ldots, u_K\}$ be the user set, and

Table 1
Features extracted based on intentions of fake news.

Intention	Feature	Example	Reference
Mislead the Public	Special Symbol Features	The number of URLs The fraction of messages containing a URL Whether a message contains a personal pronoun in 1st, 2nd, or 3rd person Whether a message contains a question mark or exclamation mark The number of "@" tags in a message	[3,6,7,9–11]
Manipulate Opinions	Sentiment Features	The numbers of positive and negative emoticons used in a message Average sentiment score of a message Whether a message contains strong negative words Fraction of messages containing negative sentiment and positive sentiment	[10,13–16,20,55]
	Style Features	Style similarity of news messages	[56]
Attract User Attention	Topic Features	The fraction of hashtags (#) LDA-based topic distribution of a message	[6,13]
	Visual Features	Whether a message contains images and videos Whether a user has a profile image The time delay of an image Image distribution: clarity score, coherence score, etc. Whether an image matches the text The topic of an image The semantic of an image extracted by pretrained models	[10,17,28–30,55,57–60]
	Clickbait Features	Similarity between the headline and top sentences Informality and readability of a message Whether a message contains internet slang or swear words Whether a message uses repeated characters (e.g., ooh, aah, etc.)	[9,61]
GeneralFeatures	Temporal Features	The time difference between a repost and the original message Whether a message has periodic reposts/comments spikes The number of duplicated reposts/comments	[10,13–15]
	User Features	Number of friends, followers, and the ratio of followers and friends Whether a user is a VIP or a verified user Register time, client program type, location, organization, gender Whether social profiles in different social media are linked with each other Whether the profile of a user contains a description, URL, and location Ratio of messages containing event verbs Ratio of messages containing strong negative words The number of messages at posting time	[6–8,10,11,13,14,62–70]
	Other Linguistic Features	TF-IDF feature, part-of-speech tagging feature Bag-of-words, named entity recognizing feature	[16–20,37,39,55,57,62–67,71–75]

user information includes user profiles and social relationships. A news message m_i is originally posted by user $u_0^i \in \mathcal{U}$ at time t_0^i, and is reposted by user u_j^i at time t_j^i with textual content z_j^i about m_i. Thus, this repost is a triplet $p_j^i = \{u_j^i, t_j^i, z_j^i\}$, and all post/reposts set $\mathcal{P}_i = \{u_0^i, t_0^i, m_i\} \bigcup \{p_j^i = \{u_j^i, t_j^i, z_j^i\}\}$ of message m_i form a propagation structure. Besides, a user u_k^i comments on message m_i at time t_k^i with comment textual content τ_k^i. Let $C_i = \{c_k^i = \{u_k^i, t_k^i, \tau_k^i\}\}$ be the comment set of m_i. The goal of fake news detection is to learn a function $f(\cdot)$ using above information to distinguish whether a message m_i or an event x is fake or not.

3. Intentional creation

The first discriminative characteristic is that fake news is intentionally created. Intentional feature-based methods are employed, which first extract distinctive features according to the intention, and then perform classification using extracted features.

3.1. Feature extraction

As shown in Table 1, intentional features of fake news can be categorized into four types: misleading the public, manipulating opinions, attracting user attention and other general features. Except for the general features, the other three types exhibit different features that reflect distinct underlying intentions of their creation. In Fig. 2, we provide typical examples for each of them with corresponding feature words or symbols marked in red color, and we discuss their differences as follows.

• Mislead the Public: the intention behind this type of fake news is to make people believe that the content of the news is true. Thus, its textual content is quite similar to that of real news, except for subtle differences of employing special symbols, such as personal pronouns, URLs and "@" tags, etc, in order to make the news seem more convincing as shown in the example of Fig. 2a. Therefore such symbols are extracted as features of this type of fake news.

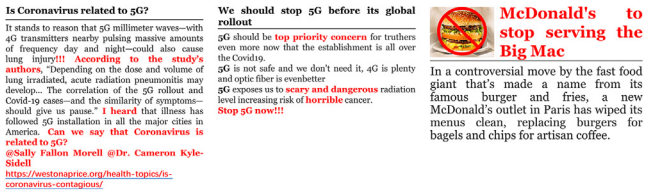


Fig. 2. Examples of fake news with three different types of creation intentions: (a) misleading the public, (b) manipulating opinions and (c) attracting user attention.

• Manipulate Opinions: The primary objective of this category of fake news is to manipulate individuals' opinions in order to gain their support for the viewpoints of fake news. Different from the above category of "Mislead the Public", this type of fake news does not focus on making the news appear more authentic, but rather aims to convey a certain viewpoint, and meanwhile uses emotional words and particular writing styles to incite people to support that viewpoint. As shown in the example of Fig. 2b, it incites people to support "Stop 5G". Thus, such sentiment features and style features are captured for detection of this type of fake news.

• Attract User Attention: this type of fake news aim to increase traffic, click rate or to create a buzz. Different from the above two categories, this type of fake news heavily relies on the headlines or cover page images as shown in the example of Fig. 2c, and therefore hot topics, attractive images or clickbaits are often employed. Accordingly, topic features, visual features and clickbait features are extracted for detection.

3.1.1. Mislead the public

This kind of fake news is usually used for commercial purposes, e.g., to make the veracity of a news message indistinguishable for the public so that users are more likely to buy some products. It often employs special symbols as listed in Table 1 to make it seem more convincing, which can be extracted as features.

Special Symbol Features: These features focus on capturing some special words or characters that are typically used to mislead the public. For instance, Castillo et al. [6] propose to consider the text length of news messages, and whether the message contains question or exclamation marks. This is because fake news generally has similar length and uses special marks to mislead users. Gupta et al. and Castillo et al. [3,6] propose to count the number of words, and use the first, second or third-person pronoun, as a feature. Liu et al. [7] propose to verify whether the message contains witness phrases, like "I see", "I hear" and so on. This is because a news message seems more credible if containing such words. Besides, Gupta et al. [8] and Sahoo et al. [76] consider the external URLs in messages as a supportive evidence, and Biyani et al. [9] extract some features from URLs, such as frequencies of dashes, upper case letters, and commas. Sun et al. [10] propose to consider the influence of messages, and thus the number of "@" tags, comments and reposts are calculated. Yang et al. [11] study the location feature of the event described in a message, and they show that if the location is a foreign place, the message has a higher probability to be fake than that of a domestic place.

3.1.2. Manipulate opinions

Fake news is often used to manipulate people's opinions, especially for political purpose. As discussed above, this type of fake news often employs emotional words and particular writing styles, which are captured as sentiment features and style features, respectively.

Sentiment Features: Fake news tends to use emotional or inflammatory words to manipulate users to support its viewpoints. Sentiment features [9,10,13–16,18,55,60,77] are extracted to identify such inflammatory speech that contains emotional words or sentences. To extract such sentiment features, many sentiment analysis tools are exploited. For instance, [15,16] utilize a sentiment tool referred to as the Linguistic Inquiry and Word Count (LIWC) to count the number of words in psychologically meaningful categories. On this basis, a large amount of sentiment-related statistic features are extracted. For Sina Weibo messages, Sun et al. [10] consider whether a message contains strong negative sentiment words and opinion words. Wu et al. [14] employ the number of positive or negative sentiment words within a message and compute the average sentiment score of the message. For Twitter messages, Ma et al. [13] identify positive or negative words using MPQA3 sentiment lexicon and some manually collected frequent emoticons. Sheng et al. [78] design a pattern-based model that extracts features from negation and sentiment words of Weibo and Twitter messages for veracity verification.

Differently, some works find that fake news prefers to use some emotional extreme adverbs or adjectives. Hence, the Named Entity Recognition [60] and Parts of Speech (POS) [18,60] techniques can be employed. To be specific, Hassan et al. [18] exploit the Natural Language Toolkit (NLTK) tagger to extract the POS features. They collect 43 POS tags in the corpus, and count the number of words belonging to these tags for each sentence.

Style Features: Fake news is generally written with distinctive styles in order to manipulate opinions. Style features are employed to model writing styles of messages. In particular, political hyperpartisan news is more likely to be fake news since it tend to manipulate user opinions, and its writing style is different from mainstream news. Potthast et al. [56] analyze the writing style of hyperpartisan news, and reveal that the style of left-wing and right-wing news is similar to each other, but different from mainstream news. Koppel et al. [79] propose an unmasking scheme to separate hyperpartisan news from mainstream real news, which can be used to detect fake news with the intention to manipulate opinions. Zhu et al. [80] examine the differences in writing styles between real and fake news articles from eight distinct perspectives, including readability, logic, credibility, formality, interactivity, interestingness, sensationalism, and integrity.

3.1.3. Attract user attention

This kind of fake news is mainly used for commercial or entertainment purposes, e.g., to increase traffic, click rate or to create a buzz, where topic features, visual features and clickbait features are extracted to differentiate fake news from real news.

Topic Features: Fake news tends to make use of sensational topics to attract users' interest [6], such as divorces or pregnancies of celebrities and flight accidents (e.g. "Flight MH370 lost contact"). Based on this observation, the topics a message relates to can be used as one type of features for detecting fake news. For example, Ma et al. [13] adopt the topic distribution of a message as the feature, which is calculated with the Latent Dirichlet Allocation (LDA) model [81].

Visual Features: Fake news tends to associate images or videos with the message as visual descriptions, which are much more eye-catching than textual content [59]. Given that forged images/videos [82,83] (produced by manipulation operations such as splicing, copy-move or retouching, or even created by computers) are sometimes attached with fake news, it is intuitive to use forensic based methods [84–86] to detect such multimedia data. However, these methods might be ineffective in social web [87], since the forged data usually undergo multiple post-processing operations, such as recompression and filtering during dissemination, which to some extent destroys the forensic traces. In this paper we focus on visual features particularly for fake news detection, which are categorized as visual statistical features and visual semantic features. The former focuses on statistical distributions and the latter focuses on semantics of visual contents.

Visual statistical features can be extracted from associated images or videos to detect fake news. For example, Sun et al. [10] propose to detect outdated images as a clue of fake news. They start a query for the image using an image search engine to retrieve all the records from the internet, sort the search results chronologically, and the eldest entry gives the original publish time of this image. If the time span (i.e. the time difference between the posting time of the news and the original publish time of this image) is bigger than a predefined threshold, the image is considered as outdated, and the corresponding message is more likely to be fake news.

Jin et al. [59] propose five visual statistical features to measure image distribution: visual clarity score, visual coherence score, visual similarity distribution histogram, visual diversity score, and visual clustering score. In this work, related news messages about the same event are grouped together for event-level fake news detection. They observe that a real event usually has images from different sources, and its image distribution tends to be general, while a fake event usually has limited sources of images and its image distribution tends to be distinct from the average. Based on this observation, statistical scores are designed to model image distributions of events for fake news detection.

Visual semantic features aim to detect fake news by examining the coherence of visual content, textual content and event in the semantic level. In particular, fake news tends to attach pictures to increase its credibility, nevertheless such pictures are usually irrelevant with the news event as shown in [52]. To detect fake news with images, Sun et al. [10] first use the attached image as a query to retrieve similar pictures from the search engine, returning a set of websites that are ranked based on their credibility. The text messages are crawled from the top ranked web. Then, the Jaccard coefficient is calculated between the text of the news message and the text crawled above. If the value of Jaccard coefficient is low, the news message is considered as text-image unmatched fake news.

Additionally, the visual content is useful for clustering messages into groups, and the group level fake news detection can be performed. Specifically, Jin et al. [17] propose to cluster messages with the same

image or video as a group. Features of messages in the same group are aggregated for group level fake news detection. Aside from this, some works [28–30,55,58,88] propose to employ deep neural networks to extract visual semantic features. For example, Jin et al. [28] propose a multimedia fusion network, including a visual sub-network with the VGG-19 [89] as a backbone to extract 512-dimensional visual representation. The extracted visual representation is then concatenated with the textual representation for fake news detection. Furthermore, Qi et al. [58] propose to incorporate the features of frequency and pixel domains of images to detect fake news.

Clickbait Feature: Fake news tends to use sensational headlines to induce users to click on a particular web page [9], e.g. "Beers Americans No Longer Drink". This kind of news messages is less formal and more readable than professionally written ones. To detect the clickbait, Biyani et al. [9] extract statistical informality/readability features to differentiate clickbaits, such as whether containing internet slang or swear words, whether using repeated characters (e.g., ooh, aah, etc.), and the similarity between the headline and top sentences. Besides, they further design indicative scores at the informality level and readability level, which are computed as follows:

- *Coleman-Liau score (CLScore)* [90]*:* measures reading difficulty empirically, computed as:

$$CLScore = 0.0588L - 0.296S - 15.8 \tag{1}$$

where L denotes the average amount of letters per 100 words, and S denotes the average amount of sentences per 100 words.

- *RIX and LIX indices* [91]*:* indicate readability, computed as:

$$RIX = \frac{LW}{S} \text{ and } LIX = \frac{W}{S} + \frac{100LW}{W} \tag{2}$$

where W is the word count, LW is the long word (i.e. over 6 characters) count, S is the sentence count.

- *Formality measure (fmeasure)* [92]*:* measures the formality by counting different part-of-speech tags in the article, such as nouns, verbs and adjectives.

Besides the above indicative scores, the style of structuring clickbait headlines can also be used for detection. There is one style called forward-reference [93], where such headlines usually include teasers or obvious information gap between the headline and the article. For example, given a headline: "This Is the Most Horrifying Cheating Story", users might wonder what is "This", and hence click the web page. Biyani et al. [9] show that forward-reference is usually featured with demonstrative pronouns, personal pronouns, adverbs and definite articles, which can be used for clickbait detection.

3.1.4. General features

In addition to the intention-specific features mentioned above, there are still some general features applicable for all kinds of purposes. We categorize such general features into temporal features, user features and other linguistic features.

Temporal Features: After deliberately creating fake news, the spammers tend to make it popular as much as possible. Thus, the spread of fake news is different from real news. Kwon et al. [15] extract the temporal feature of the news spread process, and observe that fake news usually has multiple and periodic spikes for the number of reposts/comments, while real news typically has a single prominent spike. Similarly, Wu et al. [14] propose to compute the repost time feature, which is the average time difference between the original message and the reposts. Considering that spammers repeatedly repost and comment with similar content, Sun et al. [10] propose to compute the number of duplicated reposts/comments as a feature. They measure the similarity between two reposts/comments by computing the Jaccard coefficient of their keywords. Reposts/comments will be considered as duplicated as long as their similarity exceeds a predefined threshold.

User Features: Users intentionally distribute fake news and may behave differently. In the literature, user features can be extracted from two perspectives: social reputation and personal information.

- *Social reputation:* Famous users are less likely to create fake news. Sometimes, spammers pretend to be famous, e.g. using nicknames similar to experts, in order to make their posts seem more credible. To conquer this, users' social engagements can be used to evaluate their social reputations. For a given user, Gupta et al. [8] propose to consider the number of friends, followers, status, and whether being verified or trusted by social media. These features indicate whether the user is popular and convincing. Sun et al. [10] propose that if a user have few followers but many followee, he is likely to be a spammer. Besides, they study the proportions of user posted messages which contain strong negative words and event-related verbs (i.e. verbs usually used for event description rather than for daily life). The larger proportion will give a higher probability for this user to be a spammer. The study conducted by [75] analyzed how users behave on Twitter when promoting false cancer treatments. They then suggested a model that focuses on users to identify those who are likely to spread health misinformation. This model extracts various user behavior features, such as attitudes, writing styles, and sentiments expressed in their posts on social media. Ghenai et al. [68], Zhao et al. [69] and Prasannakumaran et al. [70] study the features of user behavior and engagement for health misinformation dissemination, which can be exploited for health misinformation detection. For example, Ghenai et al. [68] propose a user-centric model to identify those who are likely to spread health misinformation by extracting user features, such as attitudes, writing styles, and sentiments expressed in their posts on social media.

- *Personal information:* Spammers tend to hide their real information, i.e. incomplete personal information. Gupta et al. [8] consider that the spammers might register recently, and their personal information is usually incomplete. Hence, they check the register time, personal descriptions, URLs, profile images and locations. Also, they check whether the profiles in different social media are linked with each other, since normal users always link them for convenience, while spammers do not. Liu et al. [62] indicate that such personal information is the most significant factor for early fake news detection. The consistency of tweet location, profile location and event location is also indicative [7]. Yang et al. [11] find that the client program type is particularly useful in detecting fake news on Sina Weibo, which includes PC-client program and mobile-client program. They find that if a news message refers to an event happened abroad, and the message is published from a PC-client program, it is more likely to be fake news. Dou et al. [67] propose that users' historical posts in the social media reflect their personalities, sentiments and stances, which can be used to detect fake news.

Other Linguistic Features: Linguistic features are widely used for fake news detection [16–19,32,55,62,74]. For example, Hassan et al. [17,18] propose to count TF-IDF, which is a numerical statistic that reflects the importance of each word in the sentence. This can help to analyze words that are frequently used in the fake news but rarely in the real news, such as "amazing", "poisonous" and "mortal". Chen et al. [19] also propose to extract TF-IDF. They first build a dictionary with K most frequent vocabularies with a message set, and then compute the TF-IDF for these vocabularies. Each message is encoded as a vector using the TF-IDF, and the value is 0 if a word never appears in the message. Volkova et al. [32] propose to use a pretrained model to extract the GloVe embeddings of message texts. Verónica et al. [16] derive a set of rules based on context free grammar (CFG) trees using the Stanford Parser, which comprises all the lexicalized production rules. These rules are integrated with parent and grandparent nodes, which are then encoded as TF-IDF features. Likewise, bag-of-words [72], part-of-speech tagging [18,71], and named entity recognizer [73] are also employed to analyze keywords in messages for fake new detection.

3.2. Detection methods

Based on the above extracted features, classification algorithms can be applied to detect fake news. Existing methods range from tradi-

tional machine learning approaches to recent neural network based approaches.

3.2.1. Traditional machine learning methods

After extracting features based on intentional creation, proper features and classifiers will be selected for fake news detection. The feature selection methods aim to reduce the feature dimension and retain informative features, including GINI index, information gain and random forest. For example, [6,7,18] use the GINI index to investigate the importance of features in constructing a decision tree. Castillo et al. [6] find that the sentiment features, the number of tweets, friends and re-tweet counts are prominent in fake news detection. Kwon et al. [15] use the random forest and logistic model to find informative features. Specifically, the 2-fold cross-validation is conducted repeatedly, and the features are sequentially reduced from the feature set, in order to find the most important features. Biyani et al. [9] exploit information gain to rank the features, and discard features with zero information gain. Shushkevich et al. [94] analyze the word frequencies of COVID-19 fake news, which can be used as features of such health misinformation. Given the above selected features, many machine learning methods can be used to perform fake news classification, like Decision Tree, Gradient Boosted Decision Trees (GBDT) [21], Logistic Regression, Max-Entropy classifier [20] and SVM along with different kernels.

3.2.2. Neural network based methods

Inspired by recent advances in deep learning, a lot of deep neural network based approaches are proposed. Based on the network structures, such approaches can be categorized as, Recurrent Neural Networks (RNN) based methods and Convolutional Neural Networks (CNN) based methods.

RNN-based Methods: Many RNN-based methods [19,20,31,34–37,43,44,75,95] are proposed to capture the variation of fake news over time. Ma et al. [31] propose to identify fake news using RNN, which learns long-distance dependencies of information. For a given event, all related news messages are split into groups based on time intervals, and the top-K TF-IDF values of the vocabulary terms in a group are calculated as the input of each RNN unit. The final output of RNN is used for fake news classification. Rashkin et al. [20] propose a Long Short Term Memory networks (LSTM) model that takes the sequence of words as input, and predicts the reliability of the news into different categories, i.e. trusted, satire, hoax, or propaganda. Ruchansky et al. [34] propose a hybrid deep model that combines the textual content, user response (i.e. reposts/comments) and source user for more accurate fake news detection. The hybrid model consists of three key modules: capture, score, and integrate. The capture module utilizes the LSTM to capture the textual and temporal pattern of user response. The score module learns the user representation and signs a score of each user. These two modules are further integrated in the third module to perform classification.

Various frameworks are proposed to classify fake news in a more fine-grained manner. Wen et al. [35] propose to utilize external cross-lingual cross-platform features extracted by gated recurrent unit (GRU), which captures the agreement of news and corresponding comments from different social media and linguistics, and then perform fake news classification. Some works focus on attending to partial distinct features for classification, e.g. emotional words and provocative sentences. To achieve this goal, the attention mechanism [19,36,37] is employed. For example, Chen et al. [19] apply the soft attention mechanism to RNN, which can simultaneously focus on particular distinct features and capture contextual variations of the message over time. Instead of attending to the key features, some works focus on attending to key sentences. De Sarkar et al. [37] propose a hierarchical attention model for satirical news detection. Specifically, this model first takes the word embedding as input of an RNN to extract the embedding of a sentence, while all sentence embeddings are attentively merged to obtain an article level embedding that is used for classification. The experiment results show that a few key sentences, especially the last sentence of an article, play more important roles in satirical news detection.

CNN-based Methods: Many CNN-based methods [38–41,62,96] are proposed to detect fake news. Yu et al. [38] split news messages about an event into chronological groups, and a representation vector can be learned for each group by paragraph vector methods [97]. The group of vectors forms a matrix as the input of a CNN, which automatically extracts local-global features and learns high-level interactions of latent features. Additionally, external sources can be mined to facilitate fake news detection. For instance, Karimi et al. [39] present a multi-source and multi-class detection model that incorporates information from different sources to increase the discrimination ability to differentiate degrees of fakeness, including True, Mostly-True, Half-True, Barely-True. In this process, the feature of the message text from each source is extracted. Then, an attention mechanism is used to fuse such features, and the result is used for multi-class classification.

Qian et al. [40] propose to utilize the historical user responses in previous articles as auxiliary information to perform fake news early detection. The overall framework consists of a User Response Generator (URG) and a Two-Level Convolutional Neural Networks (TCNN). The URG aims to learn a generative model of user response to true and fake news based on their historical responses, and the TCNN learns features of the news in both word-level and sentence-level. The two modules are finally fused to perform classification. Wang et al. [98] consider the scenario where users can report to the social media platform whether a news message is fake or not, even though such reports may be noisy. They exploit these reports as weak annotations and use CNN for feature extraction.

4. Heteromorphic transmission

The second discriminative characteristic is that fake news is heteromorphically transmitted, i.e., the propagation structure of fake news is different from that of real news. For example, the propagation tree of fake news is typically deeper and wider [5]. As for health misinformation, the propagation of health misinformation tends to form more local clusters than real information, resulting in the unique propagation patterns of health misinformation for detection [99]. Such heteromorphic propagation structure is caused by (a) the intentional nature of users during fake news propagation. For example, spammers and bots are paid to propagate fake news for commercial or political purposes. (b) the varying nature of messages during fake news propagation, e.g., comments with incendiary language or related to hot topics. Yang et al. [100] propose a subgraph reasoning mechanism for the detection of fake news. This mechanism aims to identify the most significant subgraphs within the news propagation network, as they play a crucial role in verifying the authenticity of the news.

All these not only promote the propagation, but also result in different propagation characteristics of fake news. For example, Castillo et al. [6] propose to use the maximum or average depth of the propagation graph, the degree of the root, the maximum/average degree of the graph, etc. to detect fake news. Kwon et al. [15] propose to extract features from three types of networks, such as the friendship network, the largest connected component of the friendship network, and the diffusion network (i.e. the propagation tree for a given message). They show that if the fraction of information flow from low to high-degree nodes is large or the fraction of singletons is large in the diffusion network, the message is likely to be fake.

Therefore, the propagation structure can be used to detect fake news. According to the way in which the propagation structure is constructed, existing propagation-based methods can be categorized into three groups, as shown in Fig. 3.

- **Message-based Propagation Structure:** The propagation structure is modeled as a network where the nodes correspond to messages, and the edges correspond to pairwise relations between messages. The node information is the content of the corresponding message, such as

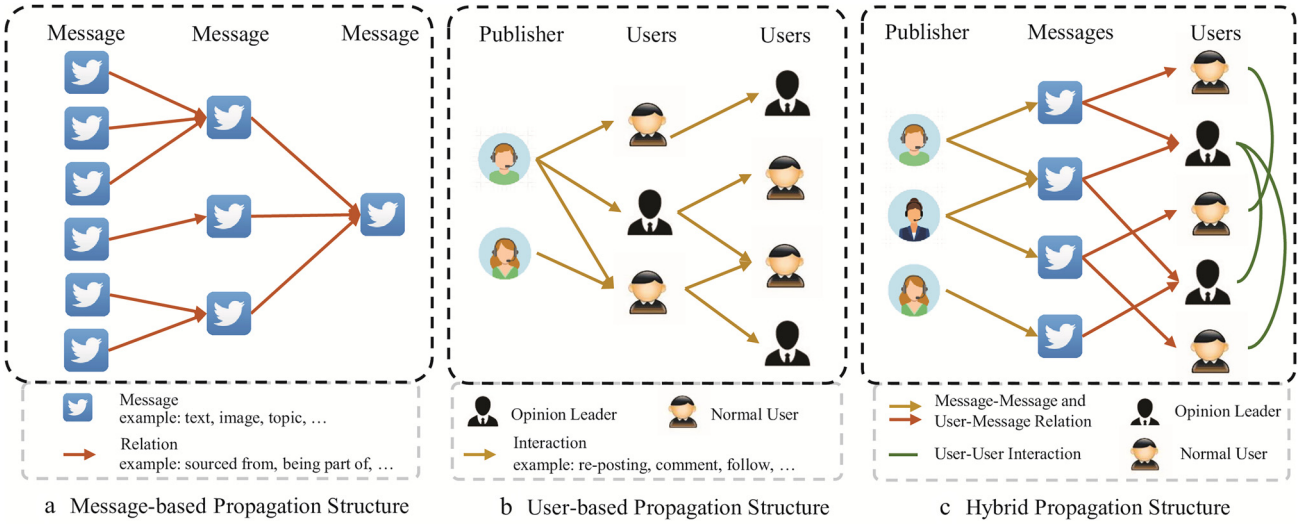


Fig. 3. According to the way in which the propagation structure is constructed, existing propagation-based methods can be categorized into three groups: (a) message-based propagation structure, (b) user-based propagation structure, and (c) hybrid propagation structure.

Table 2
Propagation based Methods for Fake News Detection.

	Characteristics	Methods	Reference
Message-based	News hierarchical propagation structure including layers, such as event layer, sub-event layer and message layer	Credit propagation mechanism for entity credibility evaluation	[12]
	Message propagation patterns/graphs	Kernel based classification method to detect fake news propagation pattern	[14,26]
		GCN based method to learn propagation and dispersion representations of the graph for detection	[74]
User-based	Topology features of propagation structures, such as the fraction of isolated nodes and the fraction of news diffusion from low degree node to high degree node	Random forest based detection method	[15]
	User propagation structure with user profiles along the propagation path, such as follower counts, verification status, etc.	RNN and CNN based method to capture the global and local features of user propagation path for detection	[102,103]
		GAT based method to learn the representation of the user propagation structure for detection	[66]
Hybrid	Event-message-user relationship with the assumption that credible entities link with each other with high weights	Credit propagation mechanism for entity credibility evaluation	[8]
	Publisher-news, news-user and user-user interactions, including news publishing, user reposting and user following etc.	Representation learning based method to learn publisher, user and news representations based on their interactions for prediction	[104]
	Propagation graph structure including nodes such as, publishers, news messages and users	CNN based method using the representations of users' profiles and their reposted texts along the propagation path as the input of the CNN classifier for detection	[62]
		GNN based method to learn the node representations in the propagation graph for detection	[63–65]

the text, image, topic, sentiment score, etc. The edge information could be different types of message relations such as sourced from, being part of, belonging to the same event and so on.

• **User-based Propagation Structure:** The propagation structure is modeled as a network, where the nodes correspond to users, and the edges correspond to pairwise interactions between users. The node information is the attributes of the corresponding user, including authenticated or not, number of followers, number of friends, etc. The edges, or user interactions, could be reposting, commenting, replying, following, unfollowing, liking, and so on.

• **Hybrid Propagation Structure:** The propagation structure is modeled as a network, where the nodes correspond to messages or users, and the edges correspond to the relations or interactions between two nodes. Typically, the relation between a user and a message is reposting or commenting on. In this kind of methods, user-message, user-user and message-message relations are fully exploited. By integrating these various relations, the credibility of users and messages can be predicted in a mutually reinforced manner. This kind of methods has become popular in recent years.

Based on the above three propagation structures, we summarize existing works in three categories, as shown in Table 2. We then discuss these mechanisms respectively.

4.1. *Message-based propagation structure*

Many traditional machine learning based methods are proposed to model message-based propagation structure. The most common strategy is kernel-based approaches, which capture the propagation patterns and differentiate fake news from real news. Wu et al. [14] propose a random walk graph kernel and a normal radial basis function (RBF) kernel to capture the high-order propagation patterns of the message. In this process, the random walk graph kernel is used for measuring the similarity between different propagation trees, and the RBF is applied to measure the distance between features of different messages. Similarly, Ma et al. [26] propose a propagation tree kernel to detect fake news by comparing the similarity between the propagation trees of different news.

Besides, some works model the propagation structure in a fine-grained level. Jin et al. [12] propose a hierarchical model that consists of three layers, i.e. event, sub-event and message layers. For a given news event, in a bottom-up manner, all related messages are first clustered into sub-events with the single-pass incremental clustering algorithm, and then all sub-events link to the news event, each of which reflects one point of view of this event. Each entity of the network is assigned with a credibility value, which will propagate through the network. An iterative graph-based optimization algorithm is proposed to calculate the final event credibility.

In recent years, a lot of research efforts has been made in graph neural networks (GNN) [101], which can capture information diffusion in a graph, and learn the high-level representations of entities in networks for downstream applications/tasks. Thus, GNN is also suitable to model the news propagation structure in social media. Bian et al. [74] propose a Bi-directional GNN, which incorporates both propagation and dispersion patterns of rumors. This model can extract features on both top-down and bottom-up paths. The top-down directed graph is leveraged to learn propagation patterns of fake news, and the bottom-up directed graph is exploited to learn dispersion patterns. Then, the learned representations are pooled and merged through multiple fully connected layers to make predictions.

4.2. User-based propagation structure

A few works focus on modeling propagation structure based on users and their interactions, which can be effective in early detection of fake news. Lin et al. [102] model each user's characteristics as a vector of values, including the length of user name, follower count, registration age, etc. For a given news message, its propagation path is presented as the sequence of users who repost it at different time instances. Thus, such path can be modeled as a sequence of user vectors, which are then fed to both an RNN-based model and a CNN-based model to capture the global and local features of the propagation path. The learned features are then concatenated for classification. Similarly, Wu et al. [103] propose a TraceMiner model that also takes the user repost sequence as the propagation structure of a message. The embeddings of users are inferred from social network structures, and LSTM-RNN is employed to model propagation structures of fake news.

Ni et al. [66] propose a MVAN model that uses both the news semantic features and the user-based propagation structure features for fake news detection. The textual content of the news is encoded by a Bi-GRU network, while the user propagation structure is modeled as a graph attention network (GAT). Their experiment results show that the key users in the propagation structure of fake news usually register lately, with no certification, nearly empty profile and very few followers.

4.3. Hybrid propagation structure

The hybrid propagation structure based approaches jointly consider users, messages and their interactions for fake news detection. For example, Gupta et al. [8] study the user-message-event structure, and propose BasicCA, which is a credibility propagation method, using a PageRank-like approach. In this model, three types of relationships are considered: (1) User-Message relationship: it is more possible for credible users to provide credible messages. (2) Message-Event relationship: the average credibility of messages associated with credible events should be higher than that with non-credible events. (3) Event-Event relationship: events that share a large number of common words and topics should obtain similar credibility. Based on this model, the credibility of users, messages and events can be jointly evaluated.

Based on the social and psychological studies that reveal the confirmation bias effect and echo chamber effect in social media, Shu et al. [104] propose a TriFN framework to exploit the tri-relationship among news publishers, news, and users, which reflects their direct interactions, such as news publishing, user reposting and user following. With the observation that publisher partisan bias is correlated with news veracity, and users tend to connect with like-minded peers, and repost news that confirms with their existing perceptions, the TriFN framework is designed to learn publisher, user and news representations based on such interactions, which can be used for fake news detection.

Liu et al. [62] propose to incorporate the repost content along with the user propagation path for early detection of fake news. They propose a neural network based method, where a Text-CNN block and a user profile embedding block are used to extract the text and user features, respectively. The vector representations of users and repost texts along the propagation path form a matrix, which is then fed to a CNN classifier for fake news detection. In addition, a PU-Learning mechanism is exploited to address the unlabeled and imbalanced training data problem.

Huang et al. [63] propose to incorporate both the user interactions and the news propagation structure for fake news detection. On one hand, users who post and repost the same news message are considered to be fully connected in an undirected graph, and GNN is used to encode user representations. On the other hand, the reposts along the news propagation path are encoded by Recursive Neural Network (RvNN). Finally, the encoded results of two networks are fused for classification. Similarly, Nguyen et al. [64] propose a Factual News Graph (FANG), which models social interactions such as user following, news posting, reposting, and source media hyperlinking as edges in the graph, and learns the representation of the target news message for fake news detection.

For a given news message, Lu et al. [65] exploit its message text, user propagation sequence, and users' profiles to verify its veracity. They propose a Graph-aware Co-Attention Network (GCAN) model, using both GRU and GNN to learn the representations of the message text and users, respectively. An attentive mechanism is jointly designed for reasonable explanations. The experiment results show that some evidential words (such as "breaking", "strict") and user profile factors (such as account creation time and user description length) have higher attentive weights for fake news detection. Yu et al. [105] propose to construct a heterogeneous graph to capture relationships between source posts, comments, and users, and employ attention-based mechanism to aggregate multi-type information for news verification.

In terms of health misinformation detection, Min et al. [106] formulate the misinformation detection as a graph classification task and model the message-message, user-user and message-user interactions on social network as a heterogeneous graph. Cui et al. [107] first extract the context information, such as news publishers and engaged users, and temporal information of user interactions, and then formulate such information in the meta paths of the propagation structure for misinformation detection. Paraschiv et al. [108] propose an approach that combines user-based, network-based, and content-based features of health misinformation with a unified meta-graph structure.

5. Controversial reception

The third discriminative characteristic of fake news is controversial reception. In social media, users have different viewpoints and comments on an event or a news message, such as supporting, denying or questioning. As discussed in [75], users tend to hold opposed stances rather than the same stance towards fake news, which are crucial for fake news detection. Therefore, recent works propose stance-based approaches that utilize the user viewpoints towards a news message to infer its veracity. User stance can be summarized as two types: explicit stance and implicit stance, as shown in Table 3.

• **Explicit Stance Based Method** utilizes the user stance as an explicit label for fake news detection. The stance label can be either from external annotations or statistical data, like the number of "thumbs up" and "thumbs down".

Table 3
Stance based methods for fake news detection.

	Characteristics	Methods	Reference
Explicit stance	News propagation graph with stance label	Pattern match based method to identify the graph pattern of fake news	[112]
	Users' comments and reposts	Multi-task GRU based scheme to detect users' stances from their comments/reposts, and predict news veracity in the same framework	[75]
	Users' opinions inferred from user behaviors, such as like, comment or repost behaviors	Bayesian network based method that considers the credibility of both news messages and users as random variables, which can be evaluated by a Gibbs sampling based scheme	[24]
Implicit stance	News messages posted by users regarding the same event	Event credibility evaluation by clustering related messages into conflicting viewpoints	[115]
	Comments, reposts, and personal information of users in the news propagation graph	GNN based method that learns the representations of users' stances from their comment/reposts and personal information for fake news detection	[57,67]

• **Implicit Stance Based Method** mines users' latent stances implicitly from their reposts or comments, which are helpful to infer the credibility of messages.

5.1. Explicit stance based method

Hanselowski et al. [33] propose to detect user stances as the first step towards fake news detection in FNC-1 [109]. Nguyen et al. [64] propose to use fine-tuned pre-trained model, such as Transformers [110] or RoBERTa [111] to detect the stances of users' comments into four categories, including neutral support (with neutral sentiment), negative support (with negative sentiment), deny, and report (i.e. repost with no comment). Similarly, Wang et al. [112] adopt sentiment analysis techniques to retrieve user attitudes towards a news message, which consists of three types of labels, including SUPPORT, DENY and QUESTION. The stances of users in the propagation path form a labeled graph. By performing a graph-based pattern matching algorithm, the distinctive patterns can be found for fake news detection. To tackle with health misinformation, Hossain et al. [113] propose to detect stances of tweets regarding specific known misconceptions with BERT and NLI (natural language inference) models, and the tweets that agree with above misconceptions will be identified as misinformation.

Ma et al. [75] propose to jointly treat the stance classification and fake news detection in a multi-task learning scheme, where these two tasks can be boosted in a mutually reinforced manner. For example, in the proposed architecture, a GRU layer can be shared for both stance classification and fake news detection tasks. Specifically, comments correlated to a given news message are organized in chronological order. Each comment is represented by a vector of TF-IDF values, which is then fed to a shared GRU in sequence, and each hidden state h_t from the GRU is used to classify the corresponding message stance, while the final GRU output h_T is used for fake news detection.

Yang et al. [24] propose a generative unsupervised approach for fake news detection, where user stances can be inferred from their behaviors, such as like, comment, or repost. The credibility of the news is viewed as a latent random variable, and a Bayesian network is exploited to capture the conditional dependencies among news veracity, users' opinions and credibility. Finally, a collapsed Gibbs sampling based approach is adopted to evaluate the credibility of news messages and users at the same time, given user opinions inferred from their behaviors.

Davoudi et al. [114] propose to construct the propagation tree and the stance network for early fake news detection, where the stance network is built by analyzing the sentiments of responses associated with a news article, and responses with similar sentiment are linked in the network. Finally, features are extracted from both the propropagation tree and the stance network to detect the news veracity.

5.2. Implicit stance based method

Implicit stance based approaches focus on mining stance latent representations from users' posts/reposts/comments, which can facilitate in measuring the credibility of news messages. Usually, posts/reposts/comments with the same stance form supportive relations, and can mutually rise their credibility, while those with conflict stances form opposed relations, and will mutually weaken their credibility.

Jin et al. [115] construct a credibility network by exploiting the relation of viewpoints. Specifically, they propose to cluster user posted news messages regarding the same event into conflicting viewpoints using the k-means algorithm. Messages are linked to construct a credibility network, and the link type is either supporting or opposing based on their viewpoints. The credibility values of the messages are propagated through the graph iteratively. Mutually supporting messages can have similar credibility values while mutually opposing messages can have opposite or close to zero credibility values. The final credibility of the event can be obtained by averaging the credibility values of all related messages.

Li et al. [116] extract users' sentiment polarity, degree of skepticism, and emoji attitude towards a news message from their responses, to facilitate the fake news detection. Dou et al. [67] propose a user preference-aware mechanism to detect fake news, which implicitly mines users' personalities, sentiments, and stances from their historical posts as user preference features. Specifically, users, who repost or comment on the same news, form a propagation graph. Their historical posts in the social media are crawled for user preference feature extraction. Then, a GNN model is used to encode user preference features in the graph, and a readout function taking the mean pooling operation over all node embeddings is performed, to obtain the entire graph embedding, which is concatenated with the news textual embedding for fake news classification.

Similarly, Xie et al. [57] propose to extract the stance representations from users' replies/comments towards a news message using BERT. All these representations are encoded by a graph-based stance reasoning network, which simulates the fact that users may aggregate others' comments, which they can browse in social media, to form their own opinions. Finally, the encoded stance representation is concatenated with the textual and visual representations of the news message for classification.

6. Datasets

With the surging interest arising in fake news detection, plenty of benchmark datasets have been proposed. Most of them are collected from real-world social media, like Twitter, Facebook and Sina Weibo. DUlizia et al. [117] provide a thorough review of evaluation datasets for fake news detection. In this paper, we give a brief introduction of popular ones that are frequently used in this field. Their statistics, i.e., the numbers of messages, events and fake news, are listed in Table 4.

• Weibo1 [31] and Weibo2 [118] are frequently used Chinese fake news detection datasets. Weibo1 is collected from a Sina community management center. Each Weibo is regarded as a news message associated with a binary label, indicating whether the story is a rumor or not. This dataset contains 4664 events, where 2313 of them are fake. In this dataset, such events are associated with 3,805,656 messages and 2,746,818 users, which include the original messages as well as retweets and replied messages. This online social context information

Table 4
Statistics of the datasets.

Statistic	Messages	Events	Fake news
Weibo1 [31]	3,805,656	4,664	2,313
Weibo2 [32]	7,300	-	3,834
Twitter15 [7]	1,490	1,490	370
Twitter16 [31]	1,101,985	992	205
LIAR [121]	12,836	-	-
FNC-1 [33]	75,385	2,587	-
MediaEval [52]	15,000	11	9,000
CCMR [35]	15,629	17	9,404
FakeNewsNet [123]	23,196	23,196	5,755
Fakeddit [124]	1,063,106	-	628,501
PHEME [125]	5,802	5	1,972
FakeHealth [126]	2,296	16	763

can help in constructing the news propagation structures and inferring users' stances towards the news. Weibo2 is a cross-domain dataset which can be found at the "Internet fake news detection during the epidemic competition held by CCF Task Force on Big Data [119]. It covers eight domains including health, economy, technology, entertainment, society, military, politics, and education, and includes 7300 news articles, where 3834 of them are fake. For each news article, user comments are also included.

• Twitter15 [7] and Twitter16 [31] are the most frequently used datasets, which are collected by Snopes [120], which is an online website providing rumor debunking service. Twitter15 contains 1490 tweets and 276,663 users, where all the tweets can be divided into four categories: non-rumors, false-rumors, true-rumors, and unverified rumors, with counts of 374, 370, 372, and 374 respectively. Twitter16 contains 992 events, including 205 non-rumors, 205 false rumors, 207 true rumors and 201 unverified rumors. Such events are associated with 1,101,985 tweets and 491,229 users.

• LIAR [121] is an American politics related fake news dataset with over 12,836 short statements during the time from 2007 to 2016. It is collected from a fact-checking website PolitiFact [122], including 4150 statements from the Democratic Party, 5687 statements from the Republican Party, 2185 statements from non-partisans, and 814 other statements. Each statement is assigned with a truthfulness rating out of 6 levels, i.e. pants-fire, false, barely true, half true, mostly true, and true. In the dataset, the proportions of the six truthfulness ratings are as follows: pants-fire 8.19%, false 19.60%, barely true 16.44%, half true 20.54%, mostly true 19.18%, and true 16.05%.

• FNC-I [33] is a dataset from an online fake news detection challenge "Fake News Challenge Stage 1 (FNC-I): Stance Detection", where the organizer of this challenge believes that understanding users' opinions and stances regarding a news message would be the first step towards fake news detection. Thus, in FNC-I they focus on stance detection, and include 75,385 (headline, document) pairs, where each pair is annotated by one of the four stance labels, i.e. agree, disagree, discuss, and unrelated, describing the stance of the document to the headline.

• MediaEval [52] is a dataset including fake news with misuse of multimedia content on Twitter. The retrieval of data was facilitated through the utilization of Topsy [127] that is a search engine and Twitter APIs. The dataset includes news messages related to 11 events, such as Hurricane Sandy, the Boston Marathon bombing, etc, and each message includes text content, attached image/video and several social contexts. Specifically, the dataset is divided into a development set and a test set. The development set includes 5008 true news messages with 176 authentic images, which are posted by 4756 users, and 7032 fake news messages with 185 misused images, which are posted by 6769 users. The test dataset consists of 1217 true news messages with 17 authentic images, which are posted by 1139 users, and 2564 fake news with 33 misused images and 2 misused videos, which are posted by 2447 users.

• CCMR [35] is a cross-lingual cross-platform multimedia rumor verification dataset, which extends MediaEval [52] from two perspectives. On one hand, MediaEval includes only 11 events, which are extended to 17 events in CCMR. On the other hand, news messages in MediEval are crawled from Twitter only, while CCMR also includes webpages collected from different search engines. Specifically, CCMR consists of three sub-datasets including CCMR Twitter, CCMR Google, and CCMR Baidu, which are all related to 17 events. CCMR Twitter includes a total of 15,629 tweets, with 6225 of them classified as true and 9404 of them classified as fake. CCMR Google has 4625 Google webpages, out of which 3197 are true, 729 are fake and 699 are unverified. CCMR Baidu consists of 2506 Baidu webpages, with 1393 true news, 508 fake news, and 605 unverified news.

• FakeNewsNet [123] mainly collects data from two well-known platforms with fact-checking: PolitiFact [122] and GossipCop [128] In PolitiFact, news messages are fact-checked by journalists and domain experts, and FakeNewsNet gathers 432 fake messages and 624 real messages from PolitiFact. In GossipCop, each news message is given a rating score out of 10, where messages with scores below 5 are considered as fake news, and 5323 fake messages and 16,817 real messages are collected for FakeNewsNet. Thus, FakeNewsNet contains 5755 fake news messages and 17,441 real news messages in total. In addition, FakeNewsNet includes 3 kinds of information i.e. news content with labels, social context information, and spatio-temporal information. News content information contains the news messages with text and images, and labels indicating their veracity. Social context information includes user engagements such as posting, forwarding, replies, and likes, etc. Spatiotemporal information provides the locations of users and news articles, and also timestamps of news publishment and users' responses.

• Fakeddit [124] is collected from Reddit [129], which is a well-known social media platform and online community for users to share information and discuss with each other out of interests. Fakeddit contains more than 1 million message samples of 22 different topics ranging from political news stories to simple everyday posts, where 628,501 of them are fake samples and 527,049 of them are true samples. The dataset is gathered from over 300,000 users, covering the period from March 19, 2008, to October 24, 2019. Each message includes its title, images, score, user comments, and up-vote to down-vote ratio, etc.

• PHEME [125] is collected from Twitter, regarding five newsworthy events, including Ferguson unrest, Ottawa shooting, Sydney siege, Charlie Hebdo shooting and Germanwings plane crash. It mainly samples tweets that triggered a large number of retweets, and collects 5802 such tweets in total. For each event, all of its tweets are organized in a timeline, where professional journalists and experts are assigned to review this timeline and indicate whether such tweets are rumors or not. Among all the tweets, 1972 are classified as rumors and 3830 are classified as non-rumors.

• FakeHealth [126] is collected from HealthNewsReview.org [130], which is a fact-checking project for health related news. It is a fake health news dataset with 2296 news about 16 health topics, including Cancer, Surgery and Nutrition etc. It contains 4 types of information: news contents, news reviews, social engagements and user networks. News contents include text, images, URLs, etc. News reviews consist of ratings, tags, categories and other elements. Social engagements provide tweets, replies and retweets. User networks provide profiles, timelines, followees and followers of users. It collects news messages from 2 types of information sources, i.e. news media and institutes, respectively, which correspondingly forms two subsets, HealthStory and HealthRelease within FakeHealth. HealthStory has 1218 real news messages and 472 fake news messages, while HealthRelease has 315 real news messages and 291 fake news messages.

7. Discussion and future directions

In this survey, we analyze the entire diffusion process of fake news, and summarize its three main characteristics, i.e., intentional creation, heteromorphic transmission, and controversial reception. In this section,

Table 5
Characteristics and proposed methods for fake news detection.

			Characteristics adopted in proposed methods																	
Intentional Creation	Mislead Public	Special Symbol	√	√			√													
	Manipulate Opinions	Sentiment		√	√	√								√						√
		Style						√												
	Attract User Attention	Topic	√	√	√	√														
		Visual		√										√					√	
		Clickbait					√													
	General Features	Temporal		√	√	√														√
		User	√	√	√	√			√				√		√	√	√	√		
		Linguistic							√	√	√	√	√	√	√	√	√	√	√	
Heteromorphic Transmission	Message-based		√			√			√											
	User-based															√	√	√		
	Hybrid												√		√					√
Controversial Reception	Explicit Stance									√					√					√
	Implicit Stance																	√	√	
References			[6]	[10]	[13]	[14]	[9]	[56]	[34]	[75]	[37]	[39]	[63]	[55]	[64]	[65]	[66]	[67]	[57]	[114]
Year			2011	2013	2015	2015	2016	2017	2017	2018	2018	2018	2019	2019	2020	2020	2021	2021	2021	2022

we take a close look of fake news detection methods using above characteristics in the past decade and reveal the trends of technological advances. This will provide some insights for designing effective detection methods. We then discuss remaining challenges and other directions for future study, such as robust fake news detection, impacts of LLM in fake news detection and early fake news detection, etc.

7.1. Trends of fake news detection methods

Table 5 shows the typical detection mechanisms using different characteristics in the past decade in chronological order, from which we can observe the trends of technological advances as follows:

Characteristic Selection: Characteristics from all three categories of the diffusion process tend to be utilized together for fake news detection. Before 2015, the detection mechanisms mainly use hand-crafted features in the category of intentional creation, and the topology features of the propagation structures in the category of heteromorphic transmission. Afterwards, the stance-based features are adopted and the message/user propagation-based mechanisms are proposed. Very recently, Nguyen and Dou [64,67] utilize the characteristics from all three categories of the diffusion process, including news content, user and publisher profiles, propagation structures and feedback stances. Furthermore, they incorporate all characteristics together to improve the detection performance.

Framework Design: A framework can be designed to capture the characteristics from all three categories for effective fake news detection. From Table 3, in early 2010s, traditional machine learning based methods are proposed to capture the hand-crafted features, such as [6,10]. Later on, with the development of deep neural network, CNN and RNN based methods [34,37,39] are proposed to capture the textual, visual and temporal features. These features mainly belong to the categories of intentional creation and heteromorphic transmission. Very recently, GNN based approaches [63,64,66,67] are proposed to capture the characteristics of propagation structures and users' viewpoint interactions, which belong to the categories of heteromorphic transmission and controversial reception. Furthermore, different types of neural networks can be flexibly combined in one framework to capture different types of characteristics for detection, which is illustrated to be comprehensive and effective. For example, Dou et al. [67] propose a framework to incorporate BERT and GNN, where BERT is used to extract features of news messages, user preferences and feedbacks, while GNN is used to extract the representations of news propagation structures. Finally, all features are fused in this framework for fake news detection.

Result Explanation: Detection results can be explained in a more fine-grained manner, which reveals the key factors of fake news. In early 2010s, GINI index and information gain based schemes are used to rank different hand-crafted features (e.g. [6,9]), and this explains which features are more important. With the recent advance in explainable deep neural networks, such detection mechanisms can reveal the key factors of fake news in a more fine-grained manner. For example, the works in [65,66] show that some evidence words, sentences and users in the propagation structures play more important roles in fake news detection.

The above discussions provide insights of designing effective and explainable fake news detection mechanisms by (1) utilizing characteristics from all three categories of the diffusion process; (2) capturing these characteristics in one framework for effective detection; and (3) making use of explainable schemes.

7.2. Future directions

Although a lot of progress has been made in the past decade, there are still remaining challenges requiring future research. In this paper, we discuss potential directions, which we believe are important and urgent.

Robust Detection: Robustness in fake news detection encompasses the ability of accurately identifying fake news and maintaining trustworthiness when facing interference or adversarial attacks. This research direction has gradually gained attention and exploration in recent years, and research efforts have been made mainly in two perspectives. On one hand, adversarial attack methods are proposed to undermine the effectiveness of fake news detectors. On the other hand, some works investigate which features and properties in current detection methods exhibit better resistance or robustness against attacks.

- Adversarial attack methods: The works in [131–136] focus on attacking from the perspective of fake news contents. For instance, the works in [131–133] examine the robustness of fake news detectors by conducting attacks that distort news content or inject adversarial words. Wang et al. [137] propose to simulate the adversarial behaviors of fraudsters and affect the feature of propagation so as to attack GNN-based misinformation detectors. These mechanisms show that attacks on detection methods can effectively reduce their performance, and demonstrate the vulnerabilities of current fake news detection methods. Therefore it is urgent to develop robust fake news detection.
- Robust properties of fake news detectors: Mahabub et al. [138] apply an ensemble voting classifier based on various machine learning algorithms and prove its effectiveness. Horne et al. [134] show that handmade content-based features, e.g. writing style, are rather robust to changes in the news cycle. Ali et al. [132] evaluate the performance of fake news detectors under different neural networks and configurations. They find that RNNs are more robust, and also increasing the maximum length of input sequences can improve the detectors' robustness.

Health Misinformation Detection: Health misinformation is defined as human health related misinformation that is false or inaccurate based on current scientific consensus [139]. Compared to general false information, the identification of health misinformation requires specialized biomedical knowledge. It can be challenging for the general public to distinguish. In addition, health misinformation is particularly related to human health and can potentially harm the physical well-being of deceived individuals. In recent years, with the outbreak of the COVID-19 pandemic, a significant amount of pandemic-related fake information has been spread, causing significant disruptions to epidemic prevention. Therefore, combating health misinformation has become increasingly important. In recent years, research efforts have been made to detect health misinformation mainly from two perspectives. On one hand, similar to the general fake news detection, features extracted from the misinformation diffusion process are exploited for detection, and thus such methods can be integrated into our proposed classification of fake news detection. On the other hand, since health misinformation involves biomedical knowledge, and fact-checking based on biomedical knowledge graphs is a promising research direction. The details are discussed as follows.

• Features extracted from the misinformation diffusion process: Such methods can be integrated into our proposed classification. For example, the work [68] belongs to intention creation, which proposes a user-centric model to identify those who are likely to spread health misinformation by extracting user features, such as attitudes, writing styles, and sentiments expressed in their posts on social media. Min et al. [106] formulate the detection as a heterogeneous graph classification task and model the message-message, user-user and message-user interactions on social network with a divide-and-conquer strategy, which is categorized into heteromorphic transmission. In the category of controversial reception, Hossain et al. [113] detects stances of post-response pairs with BERT and NLI (natural language inference) models to identify whether a claim contains misinformation.

• Fact-checking based on biomedical knowledge graphs: Biomedical knowledge graph can be used as an effective aid for health misinformation detection, which can identifies unreasonable relations between entities in texts to improve detection performance and provide explanations [95,140–142]. For instance, Cui et al. [95] apply a knowledge-guided graph attention network to capture crucial entities of news articles, and automatically assign more weights to important relations in differentiating misinformation from fact. Weinzierl et al. [142] design a Misinformation Knowledge Graph (MKG), where the misinformation detection task is formulated as a graph link prediction problem.

Impacts of LLMs in Fake News Detection: Recently, with the rapid development of LLMs such as ChatGPT, there has been a significant impact on the field of fake news detection. On the positive side, LLMs can be used for fake news detection. However, on the negative side, LLMs also facilitate the generation of fake news, presenting new challenges for fake news detection. Specifically:

• On the positive side, LLMs can be used for fake news detection. LLMs such as Chatgpt are trained on a vast amount of factual data, and can help in detecting fake news by leveraging their comprehension of factual knowledge. Although as shown in [143,144], the fact-checking accuracy of LLMs still lags behind human fact-checkers in renowned organizations like PolitiFact and Snopes, LLMs have demonstrated their potential in fact verification. Furthermore, with the ability to browse internet information and access up-to-date knowledge, LLMs have the potential to conduct real-time fact-checking.

• On the negative side, firstly LLMs exhibit the issue of "hallucinations", and they may produce text that contradicts facts "unintentionally". Secondly, even with safety mechanisms to mitigate potential risks of generating harmful or misinformation, "intentionally" designed prompts can easily bypass such restrictions to generate misinformation. However, currently there is very few research work on detecting fake news generated by LLMs, which is still an open issue, while there exists a considerable amount of research on a related task, i.e. machine-generated text detection, which can help in detecting news generated by LLMs. These methods can be roughly categorized into two types, including statistical feature based methods [145–147] and neural language model based methods [148–151]. The former mainly focuses on the differences in statistical features between machine-generated texts and human-written texts, while the latter extracts deep features from the text semantics to detect machine-generated texts.

Early Detection: To maximally reduce the negative impact of fake news, it is crucial to detect fake news as early as possible, before it is widely spread. However, it is challenging to detect fake news early, since the available information is limited, where basically only the news content, user profiles and propagation information at the early stage can be used. Though the works in [19,62] have explored the early detection, the performance still needs to be improved. To address this issue, similar to [67], besides user profiles, their historical posts/reposts/comments can be mined to infer their personalities, sentiments and stances, which provides extended user information for early fake news detection. In addition, the study of users' historical behaviors can help infer whether a user is a spammer. If spammers can be marked beforehand, it will facilitate early fake news detection and prevent fake news propagation.

Unsupervised/Semi-supervised Approaches: Supervised machine learning based methods prevail in fake news detection, which usually rely on a large amount of hand-labeled training data. However, labeling such data consumes a lot of time and efforts. Furthermore, the labeled data may soon be outdated given the dynamic nature of social media and news [98]. Thus, unsupervised or semi-supervised approaches should be explored to tackle such issue.

Similar to Jin et al. [115], clustering based unsupervised methods can be used to cluster news messages or users' feedbacks, and such information can be further used to estimate the veracity of news messages. Besides, a semi-supervised model with partially labeled data can also be exploited. Similar to Liu and Wu [62] and Wang et al. [98], partially labeled data can be used to generate useful annotations for unlabeled data in a reinforcement manner, which can enlarge the training set and improve the detection performance.

Fact-Checking based Approaches: In this survey, we focus on the detection methods that use the characteristics and features of news contents and related social contexts, while there is another branch of methods referred to as fact-checking based methods, which check the veracity of news with the ground truths. For example, there are many websites providing fact-checking service, such as Factchecker [152], PolitiFact [122], etc. They all rely on expert manual fact-checking, which is not able to scale with the rapid increase of news spread in social media. To address this issue, automatic fact-checking algorithms are studied. Shi et al. [22] and Wu et al. [23] perform fact-checking based on knowledge graph. For example, Shi et al. [22] view the fact-checking as a link-prediction task in the knowledge graph extracted from Wikipedia and SemMedDB. They perform a DFS-like graph traversal algorithm to retrieve meta paths, and the top-k discriminative paths are extracted as features to train the logistic regression model. Ciampaglia et al. [153] propose a semantic proximity metric that performs fact-checking by finding the shortest path between concept nodes on knowledge graphs. Generally, if the news content falls in the range of the knowledge domain, this type of methods may detect fake news with high accuracy. However, a lot of news regarding new events is published and spread through social media every day. This requires the fact database or knowledge graphs to be expanded and updated frequently, which may also raise open issues in this field.

8. Conclusion

In this paper, we provide a thorough survey of fake news detection techniques from a brand-new perspective, i.e. the intrinsic characteristics of fake news diffusion process, including intentional cre-

ation, heteromorphic transmission and controversial reception. This review can not only guide researchers to better understand this field, but also help us reveal the trends of technological advances, which provides insights on how to design effective and explainable fake news detection mechanisms. We also discuss popular benchmark datasets and further suggest several future directions in fake news detection.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China Science Fund for Creative Research Groups (62121002) and Excellent Young Scientists Fund (62222212).

References

[1] Z. Jin, J. Cao, H. Guo, et al., 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, 2017, pp. 14–24.
[2] M. Takayasu, K. Sato, Y. Sano, et al., Rumor diffusion and convergence during the 3.11 earthquake: A twitter case study, PLoS One 10 (4) (2015) e0121443.
[3] A. Gupta, H. Lamba, P. Kumaraguru, et al., Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 729–736.
[4] F. Alam, F. Dalvi, S. Shaar, et al., Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, 2021, pp. 913–922.
[5] X. Zhou, R. Zafarani, Fake news: A survey of research, detection methods, and opportunities, arXiv preprint arXiv:1812.00315 2 (2018).
[6] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.
[7] X. Liu, A. Nourbakhsh, Q. Li, et al., Real-time rumor debunking on twitter, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015, pp. 1867–1870.
[8] M. Gupta, P. Zhao, J. Han, Evaluating event credibility on twitter, in: Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM, 2012, pp. 153–164.
[9] P. Biyani, K. Tsioutsiouliklis, J. Blackmer, "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016.
[10] S. Sun, H. Liu, J. He, et al., Detecting event rumors on Sina Weibo automatically, in: Asia-Pacific Web Conference, Springer, 2013, pp. 120–131.
[11] F. Yang, Y. Liu, X. Yu, et al., Automatic detection of rumor on Sina Weibo, in: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012, pp. 1–7.
[12] Z. Jin, J. Cao, Y.-G. Jiang, et al., News credibility evaluation on microblog with a hierarchical propagation model, in: 2014 IEEE 14th International Conference on Data Mining, IEEE, 2014, pp. 230–239.
[13] J. Ma, W. Gao, Z. Wei, et al., Detect rumors using time series of social context information on microblogging websites, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015, pp. 1751–1754.
[14] K. Wu, S. Yang, K.Q. Zhu, False rumors detection on Sina Weibo by propagation structures, in: 2015 IEEE 31st International Conference on Data Engineering, IEEE, 2015, pp. 651–662.
[15] S. Kwon, M. Cha, K. Jung, et al., Prominent features of rumor propagation in online social media, in: 2013 IEEE 13th International Conference on Data Mining, IEEE, 2013, pp. 1103–1108.
[16] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, et al., Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3391–3401.
[17] Z. Jin, J. Cao, Y. Zhang, et al., MCG-ICT at MediaEval 2015: Verifying multimedia use with a two-level classification model, in: Working Notes Proceedings of the MediaEval 2015 Workshop, 2015.
[18] N. Hassan, C. Li, M. Tremayne, Detecting check-worthy factual claims in presidential debates, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015, pp. 1835–1838.
[19] T. Chen, X. Li, H. Yin, et al., Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 40–52.
[20] H. Rashkin, E. Choi, J.Y. Jang, et al., 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, 2017, pp. 2931–2937.
[21] X. Wang, C. Yu, S. Baumgartner, et al., Relevant document discovery for fact-checking articles, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 525–533.
[22] B. Shi, T. Weninger, Fact checking in heterogeneous information networks, in: Proceedings of the 25th International Conference Companion on World Wide Web, 2016, pp. 101–102.
[23] Y. Wu, P.K. Agarwal, C. Li, et al., Toward computational fact-checking, Proc. VLDB Endow. 7 (7) (2014) 589–600.
[24] S. Yang, K. Shu, S. Wang, et al., Unsupervised fake news detection on social media: A generative approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 5644–5651.
[25] S. Hosseinimotlagh, E.E. Papalexakis, Unsupervised content-based identification of fake news articles with tensor decomposition ensembles, in: Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web, 2018.
[26] J. Ma, W. Gao, K.-F. Wong, Detect rumors in microblog posts using propagation structure via kernel learning, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 708–717.
[27] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[28] Z. Jin, J. Cao, H. Guo, et al., Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 795–816.
[29] Y. Wang, F. Ma, Z. Jin, et al., EANN: Event adversarial neural networks for multi–modal fake news detection, in: Proceedings of the 24th ACM International Conference on Knowledge Discovery and Data Mining, 2018, pp. 849–857.
[30] Z. Jin, J. Cao, J. Luo et al., Image credibility analysis with effective domain transferred deep networks, arXiv preprint arXiv:1611.05328(2016).
[31] J. Ma, W. Gao, P. Mitra, et al., Detecting rumors from microblogs with recurrent neural networks, in: International Joint Conference on Artificial Intelligence, 2016, pp. 3818–3824.
[32] S. Volkova, K. Shaffer, J.Y. Jang, et al., Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 647–653.
[33] A. Hanselowski, A. PVS, B. Schiller et al., A retrospective analysis of the fake news challenge stance detection task, arXiv preprint arXiv:1806.05180(2018).
[34] N. Ruchansky, S. Seo, Y. Liu, CSI: A hybrid deep model for fake news detection, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 797–806.
[35] W. Wen, S. Su, Z. Yu, Cross-lingual cross-platform rumor verification pivoting on multimedia content, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3487–3496.
[36] K. Shu, L. Cui, S. Wang, et al., dEFEND: Explainable fake news detection, in: Proceedings of the 25th ACM International Conference on Knowledge Discovery and Data Mining, 2019.
[37] S. De Sarkar, F. Yang, A. Mukherjee, Attending sentences to detect satirical fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3371–3380.
[38] F. Yu, Q. Liu, S. Wu, et al., A convolutional approach for misinformation identification, in: International Joint Conference on Artificial Intelligence, 2017, pp. 3901–3907.
[39] H. Karimi, P. Roy, S. Saba-Sadiya, et al., Multi-source multi-class fake news detection, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1546–1557.
[40] F. Qian, C. Gong, K. Sharma, et al., Neural user response generator: Fake news detection with collective user intelligence, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, vol. 18, 2018, pp. 3834–3840.
[41] K. Popat, S. Mukherjee, A. Yates, et al., Declare: Debunking fake news and false claims using evidence-aware deep learning, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 22–32.
[42] Y. Tashtoush, B. Alrababah, O. Darwish, et al., A deep learning framework for detection of COVID-19 fake news on social media platforms, Data 7 (5) (2022) 65.
[43] S. Kumari, H.K. Reddy, C.S. Kulkarni, et al., Debunking health fake news with domain specific pre-trained model, Global Trans. Proc. 2 (2) (2021) 267–272.
[44] M.-Y. Chen, Y.-W. Lai, Using fuzzy clustering with deep learning models for detection of COVID-19 disinformation, Trans. Asian Low-Resour. Lang. Inf. Process. (2022).
[45] A. Zubiaga, A. Aker, K. Bontcheva, et al., Detection and resolution of rumours in social media: A survey, ACM Comput. Surv. 51 (2) (2018) 1–36.
[46] X. Zhang, A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Inf. Process. Manage. 57 (2) (2020) 102025.
[47] K. Shu, A. Sliva, S. Wang, et al., Fake news detection on social media: A data mining perspective, Newsl. Spec. Interest Group Knowl. Discov. Data Min. 19 (1) (2017) 22–36.
[48] D. Varshney, D.K. Vishwakarma, A review on rumour prediction and veracity assessment in online social network, Expert Syst. Appl. 168 (2021) 114208.
[49] D. Rohera, H. Shethna, K. Patel, et al., A taxonomy of fake news classification techniques: Survey and implementation aspects, IEEE Access 10 (2022) 30367–30394.
[50] I.B. Schlicht, E. Fernandez, B. Chulvi, et al., Automatic detection of health misinformation: A systematic review, J. Ambient Intell. Humanized Comput. (2023) 1–13.
[51] C. Chen, H. Wang, M. Shapiro et al., Combating health misinformation in social media: Characterization, detection, intervention, and open issues, arXiv preprint arXiv:2211.05289(2022).
[52] C. Boididou, K. Andreadou, S. Papadopoulos, et al., Verifying multimedia use at mediaeval 2015, in: Working Notes Proceedings of the MediaEval 2015 Workshop, vol. 3, 2015, p. 7.

[53] A. Bondielli, F. Marcelloni, A survey on fake news and rumour detection techniques, Inf. Sci. 497 (2019) 38–55.
[54] Q. Li, Q. Zhang, L. Si, et al., Rumor detection on social media: Datasets, methods and opportunities, in: Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, 2019, pp. 66–75.
[55] L. Cui, S. Wang, D. Lee, Same: Sentiment-aware multi-modal embedding for detecting fake news, in: Proceedings of the 2019 ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 41–48.
[56] M. Potthast, J. Kiesel, K. Reinartz, et al., A stylometric inquiry into hyperpartisan and fake news, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 231–240.
[57] J. Xie, S. Liu, R. Liu, et al., SERN: Stance extraction and reasoning network for fake news detection, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2021, pp. 2520–2524.
[58] P. Qi, J. Cao, T. Yang, et al., Exploiting multi-domain visual information for fake news detection, in: 2019 IEEE International Conference on Data Mining, IEEE, 2019, pp. 518–527.
[59] Z. Jin, J. Cao, Y. Zhang, et al., Novel visual and statistical image features for microblogs news verification, IEEE Trans. Multimed. 19 (3) (2016) 598–608.
[60] C. Boididou, S.E. Middleton, Z. Jin, et al., Verifying information with multimedia content on twitter, Multimed. Tools Appl. 77 (12) (2018) 15545–15571.
[61] Y. Chen, N.J. Conroy, V.L. Rubin, Misleading online content: recognizing clickbait as “false news”, in: Proceedings of the 2015 ACM Workshop on Multimodal Deception Detection, 2015, pp. 15–19.
[62] Y. Liu, Y.-F.B. Wu, FNED: A deep network for fake news early detection on social media, ACM Trans. Inf. Syst. 38 (3) (2020) 25.
[63] Q. Huang, C. Zhou, J. Wu, et al., Deep structure learning for rumor detection on twitter, in: International Joint Conference on Neural Networks, IEEE, 2019, pp. 1–8.
[64] V.-H. Nguyen, K. Sugiyama, P. Nakov, et al., FANG: leveraging social context for fake news detection using graph representation, in: CIKM ’20: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1165–1174.
[65] Y.-J. Lu, C.-T. Li, GCAN: graph-aware co-attention networks for explainable fake news detection on social media, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 505–514.
[66] S. Ni, J. Li, H.-Y. Kao, MVAN: multi-view attention networks for fake news detection on social media, IEEE Access 9 (2021) 106907–106917.
[67] Y. Dou, K. Shu, C. Xia, et al., User preference-aware fake news detection, in: Proceedings of the 44th International ACM Conference on Research and Development in Information Retrieval, 2021, pp. 2051–2055.
[68] A. Ghenai, Y. Mejova, Fake cures: user-centric modeling of health misinformation in social media, Proc. ACM Hum.-Comput. Interact. 2 (CSCW) (2018) 1–20.
[69] Y. Zhao, J. Da, J. Yan, Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches, Inf. Process. Manage. 58 (1) (2021) 102390.
[70] P. Dhanasekaran, H. Srinivasan, S.S. Sree, et al., SOMPS-Net: Attention based social graph framework for early detection of fake health news, in: Australasian Conference on Data Mining, Springer, 2021, pp. 165–179.
[71] A. Hassan, V. Qazvinian, D. Radev, Whats with the attitude? Identifying sentences with attitude in online discussions, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1245–1255.
[72] B. Ma, D. Lin, D. Cao, Content representation for microblog rumor detection, in: Advances in Computational Intelligence Systems, Springer, 2017, pp. 245–251.
[73] V.L. Rubin, N. Conroy, Y. Chen, et al., Fake news or truth? Using satirical cues to detect potentially misleading news, in: Proceedings of the 2nd Workshop on Computational Approaches to Deception Detection, 2016, pp. 7–17.
[74] T. Bian, X. Xiao, T. Xu, et al., Rumor detection on social media with bi-directional graph convolutional networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 549–556.
[75] J. Ma, W. Gao, K.-F. Wong, Detect rumor and stance jointly by neural multi–task learning, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 585–593.
[76] S.R. Sahoo, B.B. Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning, Appl. Soft Comput. 100 (2021) 106983.
[77] X. Zhang, J. Cao, X. Li, et al., Mining dual emotion for fake news detection, in: Proceedings of the Web Conference 2021, 2021, pp. 3465–3476.
[78] Q. Sheng, X. Zhang, J. Cao, et al., Integrating pattern-and fact-based fake news detection via model preference learning, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1640–1650.
[79] M. Koppel, J. Schler, E. Bonchek-Dokow, Measuring differentiability: Unmasking pseudonymous authors, J. Mach. Learn. Res. 8 (6) (2007) 1261–1276.
[80] Y. Zhu, Q. Sheng, J. Cao, et al., Memory-guided multi-view multi-domain fake news detection, IEEE Trans. Knowl. Data Eng. (2022) 7178–7191.
[81] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[82] L. Zhang, H. Yang, T. Qiu, et al., AP-GAN: Improving attribute preservation in video face swapping, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2021) 2226–2237.
[83] F. Peng, L. Yin, M. Long, BDC-GAN: Bidirectional conversion between computer–generated and natural facial images for anti-forensics, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 6657–6670.
[84] L. D’Amiano, D. Cozzolino, G. Poggi, et al., A patchmatch-based dense-field algorithm for video copy–move detection and localization, IEEE Trans. Circuits Syst. Video Technol. 29 (3) (2018) 669–682.
[85] S. Chen, S. Tan, B. Li, et al., Automatic detection of object-based forgery in advanced video, IEEE Trans. Circuits Syst. Video Technol. 26 (11) (2015) 2138–2151.
[86] J. Hu, X. Liao, W. Wang, et al., Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1089–1102.
[87] M. Masood, M. Nawaz, K.M. Malik, et al., Deepfakes generation and detection: STATE-OF-THE-ART, OPEN CHALLEnges, countermeasures, and way forward, Appl. Intell. 53 (4) (2023) 3974–4026.
[88] P. Xu, X. Bao, An effective strategy for multi-modal fake news detection, Multimed. Tools Appl. (2022) 1–24.
[89] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3rd International Conference on Learning Representations, 2015.
[90] M. Coleman, T.L. Liau, A computer readability formula designed for machine scoring, J. Appl. Psychol. 60 (2) (1975) 283–284.
[91] Jonathan, Anderson, Lix and Rix: Variations on a little-known readability index, Journal of Reading 26 (6) (1983) 490–496.
[92] F. Heylighen, J.-M. Dewaele, Formality of language: Definition, measurement and behavioral determinants, Interner Bericht, Center Leo Apostel,Vrije Universiteit Brüssel, 4, 1999.
[93] J.N. Blom, K.R. Hansen, Click bait: Forward-reference as lure in online news headlines, J. Pragmatics 76 (2015) 87–100.
[94] E. Shushkevich, J. Cardiff, Detecting fake news about COVID-19 on small datasets with machine learning algorithms, in: 2021 30th Conference of Open Innovations Association FRUCT, IEEE, 2021, pp. 253–258.
[95] L. Cui, H. Seo, M. Tabar, et al., DETERRENT: Knowledge guided graph attention network for detecting healthcare misinformation, in: KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2020.
[96] R.K. Kaliyar, A. Goswami, P. Narang, et al., FNDNet–a deep convolutional neural network for fake news detection, Cognit. Syst. Res. 61 (2020) 32–44.
[97] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, PMLR, 2014, pp. 1188–1196.
[98] Y. Wang, W. Yang, F. Ma, et al., Weak supervision for fake news detection via reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 516–523.
[99] L. Safarnejad, Q. Xu, Y. Ge, et al., Contrasting misinformation and real-information dissemination network structures on social media during a health emergency, Am. J. Public Health 110 (S3) (2020) S340–S347.
[100] R. Yang, X. Wang, Y. Jin, et al., Reinforcement subgraph reasoning for fake news detection, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2253–2262.
[101] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the 5th International Conference on Learning Representations, 2017.
[102] Y. Liu, Y.-F.B. Wu, Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018, pp. 354–361.
[103] L. Wu, H. Liu, 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, 2018, pp. 637–645.
[104] K. Shu, S. Wang, H. Liu, Exploiting tri-relationship for fake news detection, arXiv preprint arXiv:1712.07709 8 (2017).
[105] J. Yu, Q. Huang, X. Zhou, et al., IARNet: An information aggregating and reasoning network over heterogeneous graph for fake news detection, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–9.
[106] E. Min, Y. Rong, Y. Bian, et al., Divide-and-conquer: Post-user interaction network for fake news detection on social media, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 1148–1158.
[107] J. Cui, K. Kim, S.H. Na, et al., Meta-path-based fake news detection leveraging multi-level social context information, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 325–334.
[108] M. Paraschiv, N. Salamanos, C. Iordanou, et al., A unified graph-based approach to disinformation detection using contextual and semantic relations, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, 2022, pp. 747–758.
[109] (http://www.fakenewschallenge.org/). Last accessed December 6, 2023.
[110] J. Devlin, M.-W. Chang, K. Lee, et al., BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2018, pp. 4171–4186.
[111] Y. Liu, M. Ott, N. Goyal et al., RoBERTa: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692(2019).
[112] S. Wang, T. Terano, Detecting rumor patterns in streaming social media, in: 2015 IEEE International Conference on Big Data, IEEE, 2015, pp. 2709–2715.
[113] T. Hossain, R.L. Logan IV, A. Ugarte, et al., COVIDLies: detecting COVID-19 misinformation on social media, Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, 2020.
[114] M. Davoudi, M.R. Moosavi, M.H. Sadreddini, DSS: A hybrid deep model for fake news detection using propagation tree and stance network, Expert Syst. Appl. 198 (2022) 116635.
[115] Z. Jin, J. Cao, Y. Zhang, et al., News verification by exploiting conflicting social

viewpoints in microblogs, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.
[116] K. Li, B. Guo, J. Liu, et al., Dynamic probabilistic graphical model for progressive fake news detection on social media platform, ACM Trans. Intell. Syst. Technol. (TIST) (2022) 86.
[117] A. D'Ulizia, M.C. Caschera, F. Ferri, et al., Fake news detection: A survey of evaluation datasets, PeerJ Comput. Sci. 7 (2021) e518.
[118] Y. Wang, L. Wang, Y. Yang, et al., SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection, Expert Syst. Appl. 166 (2021) 114090.
[119] https://www.datafountain.cn/competitions/422. Last accessed December 6, 2023.
[120] https://www.snopes.com/. Last accessed December 6, 2023.
[121] W.Y. Wang, "Liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 422–426.
[122] http://www.politifact.com/. Last accessed December 6, 2023.
[123] K. Shu, D. Mahudeswaran, S. Wang, et al., FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media, Big Data 8 (3) (2020) 171–188.
[124] K. Nakamura, S. Levy, W.Y. Wang, r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, arXiv preprint arXiv:1911.03854(2019).
[125] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13–15, 2017, Proceedings, Part I 9, Springer, 2017, pp. 109–123.
[126] E. Dai, Y. Sun, S. Wang, Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 853–862.
[127] https://en.wikipedia.org/wiki/Topsy_Labs. Last accessed December 6, 2023.
[128] http://www.gossipcop.com/ Last accessed December 6, 2023.
[129] https://www.reddit.com/. Last accessed December 6, 2023.
[130] https://en.wikipedia.org/wiki/HealthNewsReview.org. Last accessed December 6, 2023.
[131] Z. Zhou, H. Guan, M.M. Bhat et al., Fake news detection via NLP is vulnerable to adversarial attacks, arXiv preprint arXiv:1901.09657(2019).
[132] H. Ali, M.S. Khan, A. AlGhadhban, et al., All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings, IEEE Access 9 (2021) 81678–81692.
[133] C. Koenders, J. Filla, N. Schneider et al., How vulnerable are automatic fake news detection methods to adversarial attacks?, arXiv preprint arXiv:2107.07970(2021).
[134] B.D. Horne, J. Nørregaard, S. Adali, Robust fake news detection over time and attack, ACM Trans. Intell. Syst. Technol. (TIST) 11 (1) (2019) 1–23.
[135] T. Le, S. Wang, D. Lee, MALCOM: Generating malicious comments to attack neural fake news detection models, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 282–291.
[136] B. He, M. Ahamad, S. Kumar, PETGEN: Personalized text generation attack on deep sequence embedding-based classification models, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 575–584.
[137] H. Wang, Y. Dou, C. Chen et al., Attacking fake news detectors via manipulating news social engagement, arXiv preprint arXiv:2302.07363(2023).
[138] A. Mahabub, A robust technique of fake news detection using ensemble voting classifier and comparison with other classifiers, SN Appl. Sci. 2 (4) (2020) 525.
[139] W.-Y. Sylvia Chou, A. Gaysynsky, J.N. Cappella, Where we go from here: Health misinformation on social media, 2020,
[140] L. Shang, Y. Zhang, Z. Yue, et al., A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media, in: 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2022, pp. 34–41.
[141] Z. Kou, L. Shang, Y. Zhang, et al., HC-COVID: A hierarchical crowdsource knowledge graph approach to explainable COVID-19 misinformation detection, Proc. ACM Hum.-Comput. Interact. 6 (GROUP) (2022) 1–25.
[142] M.A. Weinzierl, S.M. Harabagiu, Automatic detection of COVID-19 vaccine misinformation with graph link prediction, J. Biomed. Inf. 124 (2021) 103955.
[143] K.M. Caramancion, Harnessing the power of ChatGPT to decimate mis/disinformation: using ChatGPT for fake news detection, in: 2023 IEEE World AI IoT Congress (AIIoT), IEEE, 2023, pp. 0042–0046.
[144] K.M. Caramancion, News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking, arXiv preprint arXiv:2306.17176(2023b).
[145] S. Gehrmann, H. Strobelt, A.M. Rush, Gltr: Statistical detection and visualization of generated text, arXiv preprint arXiv:1906.04043(2019).
[146] G. Rosalsky, E. Peaslee, This 22-year-old is trying to save us from ChatGPT before it changes writing forever, NPR 18 (2023) 2023. Archived from the original on January
[147] E. Mitchell, Y. Lee, A. Khazatsky et al., DetectGPT: zero-shot machine-generated text detection using probability curvature, arXiv preprint arXiv:2301.11305(2023).
[148] R. Zellers, A. Holtzman, H. Rashkin, et al., Defending against neural fake news, Adv. Neural Inf. Process. Syst. 32 (2019) 9054–9065.
[149] A. Bakhtin, S. Gross, M. Ott et al., Real or fake? learning to discriminate machine from human generated text, arXiv preprint arXiv:1906.03351(2019).
[150] X. Liu, Z. Zhang, Y. Wang et al., COCO: coherence-enhanced machine-generated text detection under data limitation with contrastive learning, arXiv preprint arXiv:2212.10341(2022).
[151] W. Zhong, D. Tang, Z. Xu et al., Neural deepfake detection with factual structure of text, arXiv preprint arXiv:2010.07475(2020).
[152] https://www.factcheck.org/. Last accessed December 6, 2023.
[153] G.L. Ciampaglia, P. Shiralkar, L.M. Rocha, et al., Computational fact checking from knowledge networks, PloS One 10 (6) (2015) e0128193.

Author profile

Bo Hu received the BSc degree in Computer Science from the University of Science and Technology of China, Hefei, China in 2007, and the PhD degree in Electrical and Computer Engineering from the University of Alberta, Edmonton, AB, Canada in 2013. Currently, he is an associate professor with the School of Information Science and Technology, University of Science and Technology of China. His research interests include computational social science, recommender systems, data mining and information retrieval.

Zhendong Mao received the PhD degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, in 2014. He is currently a professor with the University of Science and Technology of China, Hefei, China. He was an assistant professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, from 2014 to 2018. His research interests include the fields of computer vision, natural language processing and cross-modal understanding.

Yongdong Zhang received the PhD degree in electronic engineering from Tianjin University, Tianjin, China, in 2002. He is currently a professor with the School of Information Science and Technology, University of Science and Technology of China. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology. He has authored over 100 refereed journal and conference papers. He serves as an Editorial Board Member of the *Multimedia Systems* Journal and the *IEEE Transactions on Multimedia*.