Artificial Intelligence Review (2024) 57:52
https://doi.org/10.1007/s10462-024-10702-9

# Disinformation detection using graph neural networks: a survey

**Batool Lakzaei**[1] · **Mostafa Haghir Chehreghani**[1] · **Alireza Bagheri**[1]

Accepted: 4 January 2024 / Published online: 14 February 2024


**Abstract**
The creation and propagation of disinformation on social media is a growing concern. The widespread dissemination of disinformation can have destructive effects on people's attitudes and behavior. So, it is essential to detect disinformation as soon as possible. Therefore, the interest in effective detection techniques has grown rapidly in recent years. Major social media and social networking sites are trying to develop robust strategies to detect disinformation and prevent its spread. Machine learning techniques and especially neural networks, have an essential role in this task. In this paper, we review different approaches for automatic disinformation detection, with a focus on methods that leverage graph neural networks (GNNs). GNNs are very suitable tools for detecting disinformation in social networks. Because on the one hand, graphs are the most comprehensive way to model social networks and on the other hand, GNNs are the best tool for processing graph data. We define different forms of disinformation, and examine the features used and the methods presented from different perspectives. We also discuss relevant research areas, open problems, and future research directions for disinformation detection in social media.

**Keywords** Social networks · Disinformation · Graph neural networks

## 1 Introduction

These days, the Internet has transformed human life by offering a fast and convenient way to access and share information. Social networks play a vital role in this process, as they allow users to connect, communicate and exchange content with each other (Zhou and Zafarani 2020). The Internet and social networks can have positive impacts, such as raising awareness on various issues (from local and global news to citizenship rights and

✉ Mostafa Haghir Chehreghani
mostafa.chehreghani@aut.ac.ir

Batool Lakzaei
b_lakzaei@aut.ac.ir

Alireza Bagheri
ar_bagheri@aut.ac.ir

[1] Department of Computer Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

environmental problems), but they also pose challenges regarding the reliability and validity of the information they disseminate. The absence of regulation and verification on social networks has enabled the rampant spread of disinformation (Zubiaga et al. 2018). Unfortunately, many users share information without verifying its source or authenticity, often driven by attractive headlines (Bondielli and Marcelloni 2019). So, fact-checking on social media plays a vital role because news and information directly or indirectly influence people's perspectives and behaviors (Shu et al. 2020).

There are several real-life examples that illustrate the potential dangers of disinformation. For instance, a false report about the bankruptcy of an airline company caused its stock value to plummet by 76%, adversely affecting its reputation and market performance (Carvalho et al. 2011). The 2016 US presidential election saw the widespread sharing of fake news on social media, with both pro-Trump and pro-Clinton stories being circulated extensively (Allcott and Gentzkow 2017). Natural disasters also serve as fertile ground for false stories aimed at creating panic and disorder such as the Japan earthquake in 2011 (Takayasu et al. 2015) and Hurricane Sandy in 2012 (Gupta et al. 2013). These examples show that the spread of disinformation can have significant consequences, impacting financial markets, political events, public security and many other fields. Therefore, it is crucial to promptly identify and counter disinformation before it spreads widely.

Initial attempts to use machine learning for disinformation detection rely on the assumption that disinformation and accurate information have distinct writing styles. However, disinformation spreaders quickly adapted writing styles, making text-based methods less effective over time and across different domains. To address this challenge, researchers explored the analysis of non-text features such as the propagation graph of information on social networks to reduce the reliance of detection models on the text. Unlike text content, the propagation graph is beyond the control of the spreaders. Therefore, propagation-based methods exhibit greater resilience against malicious actors. However, they are less effective in detecting disinformation at an early stage. As a result, recent research has sought to combine the advantages of both content-based (such as text) and context-based (such as propagation graph) methods to create more effective models for disinformation detection.

Social networks inherently form a rich and interconnected graph structure. Within this graph, nodes (representing content, users etc) are interconnected through various relationships such as citations, content similarity, reposting and friendships. The disinformation detection literature includes several graph-based methods that utilize the underlying social network over which disseminations occur, and non-graph methods that ignore the underlying graph. Non-graph methods have the tendency to treat and process information published in social networks as independent entities. Each content is analyzed in isolation, disregarding the graph structure of the social network. Therefore, non-graph methods fail to adequately capture and process these interrelated relationships.

A challenge in graph-based models is that traditional machine learning and deep learning models are not well-suited for processing graph structures, such as propagation graphs or social network graphs. The intrinsic complexities of graphs, such as non-Euclidean space and interconnected relationships, require specialized graph-based models and algorithms for efficient analysis and extraction of valuable insights. Graph neural networks (GNNs) are novel and effective neural models that are particularly developed to deal with graph data. They use a series of simple and efficient calculations, such as aggregating intermediate embeddings of nodes in the local neighborhood of a node, to compute its final embedding. In recent years, they are widely considered as one of the most effective tools for detecting disinformation on social networks. This is mainly because graphs offer a comprehensive way to model social networks, incorporating user interactions, textual and

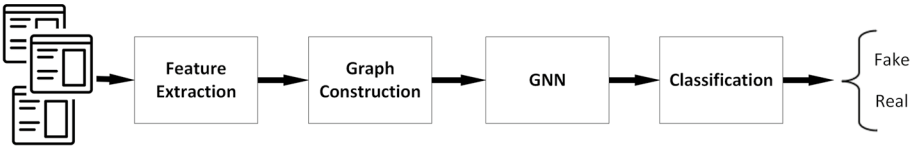


**Fig. 1** The general framework of disinformation detection using GNNs

visual content of posts, as well as explicit or implicit relationships between various elements within the network. GNNs, in turn, excel at analyzing and processing graph data, making them a natural tool for uncovering patterns and identifying disinformation within the complex network structure. The integration of graph modeling and GNNs provides a strong framework for tackling disinformation challenges.

The overall process of disinformation detection, such as fake news detection, through graph neural networks is illustrated in Fig. 1. This process can be broadly categorized into four key stages:

- Feature extraction: In this stage, initial feature vectors of data are extracted which, based on the type of information available, can be textual, visual and so on. This process utilizes diverse tools, including linguistic models and convolutional neural networks (various feature types are discussed in details in Sect. 4).
- Graph construction: In the second stage, an appropriate method for graph construction is selected, which can involve creating a similarity graph, a propagation graph, or a heterogeneous graph (we elaborate on this stage in Sect. 5.3). This process transforms the initial dataset, comprising a collection of unstructured information, into a structured graph format.
- GNN: The graph generated in the previous step undergoes processing through a GNN. This network generates an embedding vector for each node in the graph. The vector of each node captures both its structural information in its local neighborhood and its content features (we describe GNNs used for disinformation detection in section 3).
- Classification: The embedding vectors generated by the GNN serve as feature vectors and can be used by various machine learning algorithms. In the last phase, these embedding vectors are fed into an appropriate classifier, which can be a traditional machine learning algorithm or a deep neural network. The output of this classifier is the final label, which can be either binary (e.g., fake or real) or multiclass (e.g., true, false, unverified, non-rumor).

In this paper, we present a comprehensive survey of the previous and current research in disinformation detection using GNNs, which outlines existing approaches, datasets and emerging challenges. There are several surveys on disinformation detection (Asghari et al. 2022; Shu et al. 2020; Bondielli and Marcelloni 2019; Meel and Vishwakarma 2020; Shu et al. 2017; Sharma et al. 2019; Oshikawa et al. 2018; Zhou and Zafarani 2020; Saquete et al. 2020; Alzanin and Azmi 2018; Yu 2018; Mahmud et al. 2022), but very few of them concentrate on GNN-based methods (Varlamis et al. 2022; Phan et al. 2023). Our key contributions in this paper are as follows:

- We provide a comprehensive survey of disinformation, including definitions of related terms and concepts, types of features, types of approaches, the most widely used datasets and open issues.

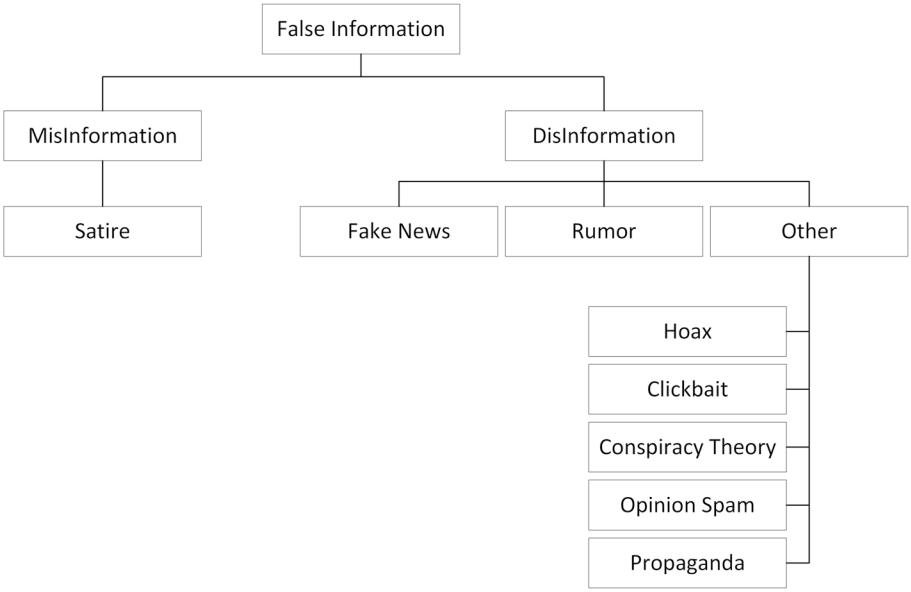


**Fig. 2** Categorization of various types of false information

- We provide a comprehensive categorization of types of features, used by disinformation detection approaches.
- For the first time, we categorize disinformation detection methods from two perspectives: features and algorithms, and describe the characteristics of each category.
- For the first time, we examine GNNs-based methods in details from the perspective of graph type and graph construction approach.

The rest of this paper is organized as follows. First in Sect. 2, we explain relevant terms widely used in the literature. Then in Sect. 3, we describe graph neural networks and their advantages. Then in Sect. 4, we examine different types of features that are extracted and used in the literature for disinformation detection. We explore different GNN-based techniques proposed recently for disinformation detection, in Sect. 5. We describe the datasets commonly used to evaluate algorithms in Sect. 6. We introduce ongoing problems and outline possible future research directions, in Sect. 7. Finally, the paper is concluded in Sect. 8.

## 2 Definitions

In this section, we explain the concepts and terms related to disinformation. In general, disinformation can take various forms, such as fake news, rumors and hoaxes. In Fig. 2, we present our categorization of different types of disinformation. Although these terms possess distinct characteristics, they are frequently used interchangeably. In the following, we provide the definitions commonly used by researchers.

*False information* content that contains false and untrue information, written for different purposes.

*Misinformation* false, mistaken, or misleading information that is often considered honest mistake. It is published as a consequence of an honest mistake, carelessness, or cognitive bias (Shu et al. 2020; Meel and Vishwakarma 2020). The purpose of misinformation is not to deceive the audience.

*Disinformation* false information that is spread deliberately to mislead and deceive (Shu et al. 2020).

*Satire* an article that incorporates cleverly crafted allusions. It does not have a harmful intention, but there is a potential for deception in it (Meel and Vishwakarma 2020).

*Fake news* news articles that are intentionally and verifiably false. It could mislead readers (Shu et al. 2020). In this definition, two crucial features are emphasized: intention and verifiability. Therefore, fake news refers to news articles intentionally created to disinform and mislead the audience, but their falseness can be corroborated by utilizing other sources (Bondielli and Marcelloni 2019).

*Rumor* information that has not yet been confirmed by official resources. It is usually spread by users (not official news resources) on social networks (Bondielli and Marcelloni 2019). This information is not necessarily incorrect, and its correctness may be confirmed in the future (Meel and Vishwakarma 2020).

*Hoax* messages that are sent to a wide range of people to persuade or manipulate them to do or prevent pre-established actions, primarily by using a threat or deception (Shu et al. 2020).

*Clickbait* using misleading headlines to encourage users to click on a particular link (Meel and Vishwakarma 2020).

*Conspiracy theory* beliefs that are primarily dismissed or disregarded by society. Conspiracy theory, on the one hand, challenges official and widely accepted explanations regarding the causes of an event. On the other hand, it attributes the event to a group of agents with illegal, hidden, and malicious intent (Shu et al. 2020).

*Opinion spam* fake or intentionally biased reviews or comments about products and services (Meel and Vishwakarma 2020).

*Propaganda* biased and misleading information that are spread in the target communities following a preconceived strategy to strengthen a particular viewpoint or particular political agenda (Meel and Vishwakarma 2020).

## 3 Graph neural networks

Graphs are a powerful method for modeling data in many real-world applications, such as the world wide web, social networks, and communication networks. Prediction and classification tasks in graph-based systems can be tackled with graph neural networks (GNNs). GNNs learn a function that maps each node to a vector in a low-dimensional vector space. This mapping should be similarity-preserving: two nodes that exhibit similar features and structural roles within the graph, should also be mapped to close points in the vector space. The vector generated for each node is called its embedding or representation vector. These embeddings can be utilized as feature vectors by various machine learning algorithms.

Figure 3 provides a high-level depiction of how graph neural networks operate. First, a neighborhood is constructed for each node. Then, in each layer, a linear transformation (trainable weight matrix) and an aggregation operation (sum, average, maximum, minimum, etc) are applied to the previous-layer embedding vectors of each node and its neighbors. These embeddings are iteratively updated and at each iteration (layer), a

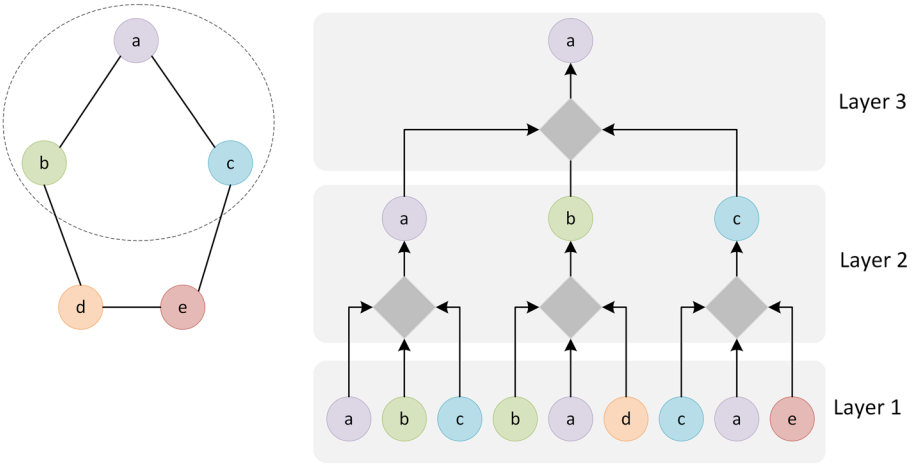


**Fig. 3** The general scheme of a graph neural network. Left: a graph (with the neighbourhood of node $a$). Right: the embedding generated for node $a$ using a 3-layer GNN. Each rhombus presents a function that consists of a linear transformation, an aggregation function (sum, mean etc), and an activation function (ReLU, sigmoid etc) (Chehreghani 2022)

new embedding is calculated for each node. In most cases, an activation function is also employed to induce nonlinearity (Chehreghani 2022). This process can be described as a message-passing mechanism (Gilmer et al. 2017). Each node follows three steps: 1) it gathers the embedding vectors of all its neighbors (message passing), 2) it applies an aggregation function to the gathered messages, and 3) it updates its own embedding vector, accordingly. Therefore, the following relation can be used to describe this mechanism (Mahmud et al. 2022):

$$h_v^l = UPDATE\big(h_v^{l-1}, AGG\big(h_u^{l-1} : u \in N(v)\big)\big), \tag{1}$$

where $h_v^l$ is the embedding of node $v$ in layer $l$, $AGG$ is an aggregation function, $UPDATE$ is a function that updates the embedding of node $v$ using its own embedding and the embeddings of its neighbors from the previous layer, and $N(v)$ is the set of nodes connected to node $v$.

### 3.1 Well-known GNN models

There exist various GNN models in the literature, that use different methods to compute and aggregate neighborhood information. Four well-known GNN models are: graph convolutional networks (GCN), graph attention networks (GAT), GraphSAGE, and graph isomorphic networks (GIN).

*Graph convolutional network*

Graph convolutional networks (GCNs) are an advancement of the convolution operator designed specifically for graph-structured data. GCN plays a pivotal role in bridging deep learning techniques with graph data, facilitating the effective integration of deep learning models into graph-based tasks. GCN employs the following function to compute node embeddings (Welling and Kipf 2016):

$$h_v^l = \sigma\left(W^{l-1} \sum_{u \in N(v) \cup \{v\}} \frac{h_u^{l-1}}{|N(v)|}\right), \tag{2}$$

where $W$ is a trainable matrix, $\sigma$ is a nonlinear function and $N(v)$ is the set of neighbors of $v$.

*Graph attention network*

The key concept behind graph attention networks (GAT) is that different neighbors of a node possess distinct roles and levels of significance. Consequently, they contribute to the generation of the node embedding with varying weights, allowing for adaptive attention-based mechanisms to capture the importance of each neighbor in the embedding process. This network employs the multi-head attention mechanism to aggregate embeddings. The following formulas show how GAT computes the embeddings (Velickovic et al. 2017):

$$h_v^l = \|_{k=1}^K \sigma\left(\sum_{u \in N(v)} \alpha_{vu}^k W^{l-1} h_u^{l-1}\right), \tag{3}$$

$$\alpha_{vu} = \sigma\big(a\big(W^{l-1} h_v^{l-1}, W^{l-1} h_u^{l-1}\big)\big), \tag{4}$$

where $\|$ is the concatenation operator, $K$ is the number of attention heads and $\alpha_{vu}$ is the attention coefficient between two nodes $v$ and $u$. The weight matrix $W$ and the parameter $a$ are learned during the training process.

In this model, the parameters $W$ and $a$, which are linear transformations, are successively multiplied by the embedding vectors. According to the principles of linear algebra, when two linear transformations are composed, the result can be represented as an equivalent single linear transformation. On the one hand, this issue can limit the expressiveness of the model. On the other hand, it makes the attention weights a monotonic function of the neighbors of a node (rather than the node itself). To address this challenge, a revised version of this model known as GATv2 has been proposed. In GATv2, attention weights are computed using the following relation (Brody et al. 2021):

$$\alpha_{vu} = a \cdot \sigma\big(\big(W^{l-1} h_v^{l-1}, W^{l-1} h_u^{l-1}\big)\big). \tag{5}$$

*GraphSAGE*

Instead of considering the entire neighborhood of each node, GraphSAGE utilizes the sampled neighborhood. This approach involves randomly selecting a subset of neighbors for each node during the aggregation step. In this way, GraphSAGE can effectively scale to large graphs while still capturing the local information necessary for node embedding generation. This model employs the following function to calculate the embeddings (Hamilton et al. 2021):

$$h_v^l = \sigma\big(W^{l-1} \ \| \ \big(h_v^{l-1}, AGG\big(h_u^{l-1}, \forall u \in N(v)\big)\big)\big), \tag{6}$$

where $\|$ is the concatenation operator and $AGG$ is an aggregation function. GraphSAGE suggests using various aggregation functions for generating embeddings. Three commonly proposed aggregation functions are average pooling, maximum pooling, and long short-term memory (LSTM):

$$AGG = \sum_{u \in N(v)} \frac{h_u^{l-1}}{|N(v)|}, \tag{7}$$

$$AGG = mean\left(MLP\left(h_u^{l-1}\right), \forall u \in N(v)\right), \tag{8}$$

$$AGG = LSTM\left(\left[h_u^{l-1}, \forall u \in (N(v))\right]\right). \tag{9}$$

*Graph isomorphic network*

In graph isomorphic networks (GIN), injective functions are employed for aggregating and updating the embeddings. GIN is developed based on the Weisfeiler-Lehman (WL) test. The WL test is a graph isomorphism test that assigns labels to nodes based on the local neighborhood structure. It checks in polynomial time whether two graphs are isomorphic or not Xu et al. (2018). GIN uses the following function to compute the embeddings (Xu et al. 2018):

$$h_v^l = MLP^l\left((1 + \epsilon^l)h_v^{l-1} + \sum_{u \in N(v)} h_u^{l-1}\right), \tag{10}$$

where $MLP^l$ is the multilayer perceptron network, used in layer $l$. Parameter $\epsilon$ can be set to a fixed scaler value, or it can be learned during the training process.

### 3.2 Advantages of GNNs

There are several key factors contributing to the success of GNNs. The first reason is their remarkable performance across a wide range of machine learning problems. GNNs have been demonstrated to be powerful and highly effective, achieving high accuracy across a wide range of problems, including classification, semi-supervised learning, and link prediction. They excel in representation learning for graph data in various domains (Guo et al. 2021; Zhou et al. 2020). Furthermore, GNNs have demonstrated their adaptability by being successfully applied to tasks that are not traditionally considered as graph problems, such as text summarization (Jiang et al. 2022; Huang et al. 2023; Luo et al. 2021) and recommender systems (Ying et al. 2018; Chen Gao et al. 2023; Wu et al. 2022). GNNs excel in modeling the connections between data, enabling them to effectively capture the item-user relations in recommender systems and sentence-sentence relations in text summarization tasks (Chehreghani 2022). This versatility showcases the broad applicability of GNNs beyond graph-specific domains, indicating their potential to extract meaningful representations from various types of data and solve diverse real-world problems.

The second reason for the success of these networks is their efficiency and scalability. A naive approach to applying neural networks to graph data is to use a graph matrix representation. However, GNNs do not rely on the matrix representation of graphs. Instead, they utilize simple and typically efficient computations, such as aggregating the local neighborhood of a node. This not only enhances their efficiency but also enables high parallelizability, allowing for faster and more scalable computations. Furthermore, various standard techniques, such as node sampling and edge removal, can be employed to further enhance the efficiency of these networks (Chehreghani 2022).

In addition, one of the less discussed advantages of GNNs is their simplicity and explainability (Chehreghani 2022). One of the limitations of many advanced machine

learning methods is the lack of a straightforward explanation for the output they produce. Indeed, while machine learning and deep learning models have made significant advancements in various domains, the lack of explainability in these models makes their functioning opaque and inscrutable to people (Dai et al. 2022; Nauta et al. 2023). This issue hinders users from comprehending, assessing, and making necessary adjustments to the system (Nauta et al. 2023). As a consequence, these methods may not be suitable for high-risk domains such as financial services, medical services, and judicial systems, where interpretability and explainability are crucial for decision-making and accountability. For such endeavors, the emerging field of eXplainable Artificial Intelligence (XAI) aims to develop models that not only exhibit comparable performance to existing efficient models but also provide outputs that are understandable and reliable for humans. XAI strives to ensure transparency in the functionality of these models, allowing users to gain insights into how decisions are made and providing explanations for the output generated.

Numerous methods have been put forth to explain the functioning of GNNs, such as GNNExplainer (Ying et al. 2019), PGExplainer (Luo et al. 2020) and XGNN (Yuan et al. 2020). What's noteworthy is that the explainability of GNNs can be likened to that of decision trees Fig. 3 (Chehreghani 2022): both of them model the learning process by constructing a rooted tree Furthermore, in both of them, computations performed in internal nodes (aggregation in graph neural networks and splitting in decision trees), are not entirely transparent. Nevertheless, there is a general belief that decision trees have higher explainability than traditional neural networks. Therefore and similarly, GNNs can be considered to be more explainable than traditional neural networks (Chehreghani 2022).

In summary, GNNs are computationally efficient and relatively explainable. They can be simplified to provide more transparent models without sacrificing performance (Chehreghani 2022). Therefore, these networks can serve as suitable models for addressing numerous challenging problems. Graphs can effectively capture and model various aspects of social networks, user interactions, textual and visual content information, as well as the explicit and implicit relationships among them. Currently, GNNs are considered to be the state-of-the-art tools for processing and analyzing graph data. Thus, GNNs can effectively contribute to the detection of disinformation within social networks.

## 4 Feature extraction

In machine learning problems, features play a crucial role as they directly impact the quality of the results. In this section, we discuss various features that have been utilized by different approaches to detect disinformation. Figure 4 presents our categorization of these features. At a high-level, they can be categorized into two main groups: content-based and context-based. In the following, we describe each category in details. We note that there exist other categorizations in the literature that can be found e.g., in Bondielli and Marcelloni (2019), Meel and Vishwakarma (2020), Shu et al. (2017), Yu (2018). In Yu (2018), the features are grouped into four categories: linguistic features, temporal properties, user information, and user interaction. In Meel and Vishwakarma (2020), the features are divided into eight types: text-specific, image-specific, user-specific, message-specific, propagation, temporal, structural, and linguistic. Our approach is similar to Bondielli and Marcelloni (2019) and Shu et al. (2017), which both distinguish between content-based and context-based features. However, unlike our approach, (Bondielli and Marcelloni 2019) does not consider visual features, and Shu et al. (2017) does not include semantic features.

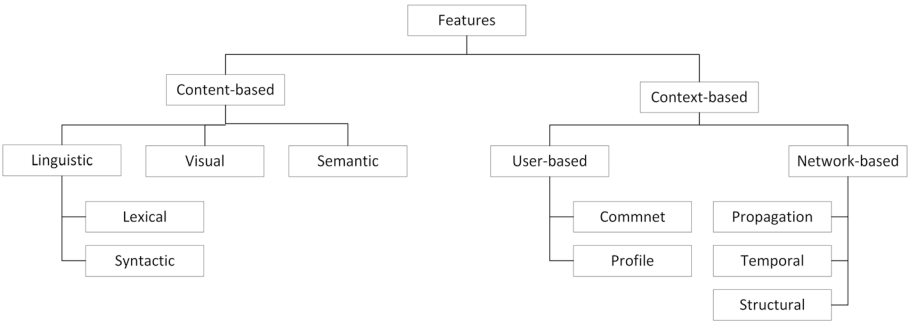


**Fig. 4** Different types of features used in the literature for disinformation detection

Moreover, neither (Bondielli and Marcelloni 2019) nor (Shu et al. 2017) provide a clear and detailed classification of user-based and network-based features, which are important aspects of the context.

### 4.1 Content-based features

Content-based features are typically extracted directly from the content. The majority of disinformation detection approaches utilize content-based features, especially textual content. Thanks to recent advances in text analysis approaches and the development of appropriate tools, extracting and accessing textual features have become increasingly effortless. However, in social networks content is not limited to text. In general, there are three types of content-based features:

- Linguistic features: Linguistic features are one of most common features used in disinformation detection (Yu 2018). Fake content is usually created with the intention to deceive audiences and serves various financial or political objectives. Therefore, it is often crafted in a manner that encourages users to share it extensively. Disinformation typically exhibits distinct writing styles and employs captivating headlines (Shu et al. 2017). Writers of disinformation employ specific phrases to evoke emotions in their audience. Common linguistic features are:
  - Lexical features: Lexical features include unigrams, bigrams and the surface forms of words in the text. One of the simplest methods is to analyze the presence of important and prominent words or content phrases (such as 2-gram and 3-gram). In addition, checking the presence of special terms, suspicious words, word length, sentence count, frequency of particular words and other factors can be effective in improving disinformation detection methods (Shu et al. 2017; Bondielli and Marcelloni 2019; Meel and Vishwakarma 2020).
  - Syntactic features: Syntactic features are related to the structure and writing style of the text. There are many syntactic features used in different methods, including the number of nouns, verbs, adverbs, and adjectives, as well as syntactic markers such as dots, question marks and exclamation marks. Other syntactic features include the usage of first-person or third-person pronouns, negative verbs and adverbs, among others (Shu et al. 2017; Bondielli and Marcelloni 2019; Meel and Vishwakarma 2020).

- Visual features: Visual features play an important role in detecting disinformation. Humans have inherent vulnerabilities and weaknesses when it comes to truth detection. So, writers often employ captivating or even fabricated images to elicit anger or other emotional responses from the audience. Visual features are extracted from visual elements of the content, such as image and video. These features include characteristics such as resolution, histogram, image ratio and object detection, among others (Shu et al. 2017).
- Semantic features: In certain methods, knowledge graphs or ontologies are employed to extract semantic features that are associated with textual or visual content. Utilizing a knowledge graph can aid in uncovering the latent semantic knowledge embedded within the content (Wang et al. 2020).

The role of content in disinformation detection is undeniably significant. So, content-based features have been widely employed in the detection of disinformation. However, these features have some disadvantages. On the one hand, a limitation of content-based features is that they are often extracted from specific domains. This restricts their generalizability since disinformation can manifest differently across various domains and involve different deceptive tactics. On the other hand, a challenge with content-based features is that writers of disinformation are aware of these features and often attempt to mimic the structure and writing style of genuine information. This makes it more difficult to differentiate between false and true information based solely on these features. In doing so, they can deceive disinformation detection models by intentionally crafting their content to closely resemble genuine information. This makes it harder for the models to accurately identify and classify disinformation based solely on content-based features. Finally, particularly in the case of rumors circulating on social media, relying solely on text-based features may not be sufficient for effective detection. This is due to the relatively short length of rumors, which may limit the amount of textual information available for analysis (Bondielli and Marcelloni 2019). As a result, alternative features and techniques, such as context-based features, need to be considered to complement the analysis and improve the accuracy of disinformation detection in such cases.

### 4.2 Context-based features

Context-based features are extracted from social networks and typically encompass the analysis of user information, sources of posts, information propagation structures, and user reactions (Bondielli and Marcelloni 2019). Generally, context-based features can be classified into two groups:

- User-based features: Disinformation is occasionally disseminated on social networks through fake accounts, such as bots. Therefore, taking into account the characteristics of users and their interactions can be beneficial in detecting disinformation. In general, user-based features represent the characteristics of users who engage with the information published on social networks (Shu et al. 2017). The most common user-based features include:
  - Profile: In social networks, users have the ability to share various information on their profiles, including a profile picture and a description of themselves. In general, various features can be extracted from the user profile, such as the number of posts, the age of the user account, the number of friends/followers and the verification sta-

tus of the user account, among others (Shu et al. 2017; Bondielli and Marcelloni 2019; Meel and Vishwakarma 2020; Yu 2018). When utilizing profile-based features, an important aspect to consider is the availability of this information. Due to privacy concerns, certain information about users and their interactions on social networks may not be accessible or available (Bondielli and Marcelloni 2019).
  - Comments: Users typically express their emotions and opinions regarding the content published on social networks through comments. Therefore, analyzing and investigating public reactions to contents can be influential in determining their authenticity. One of the simplest methods to extract features from users' opinions is to analyze sentiments and determine users' stances using text analysis and natural language processing methods (Shu et al. 2017).

- Network-based features: Social media platforms encompass various networks with diverse interests, topics and relationships. Network-based features are extracted by constructing and analyzing these special networks, such as friendship networks and propagation networks (Shu et al. 2017). These features can be categorized into the following three categories:
  - Propagation features: Studies indicate that the propagation pattern of disinformation on social networks differs from that of true information. Therefore, constructing a propagation graph/tree and analyzing its properties, such as root degree, number of nodes, average degree of nodes and tree depth can be helpful (Meel and Vishwakarma 2020).
  - Structural features: Investigating the structure of the information dissemination network and extracting structural features such as node degree and clustering coefficient, are commonly used for detecting disinformation (Shu et al. 2017; Bondielli and Marcelloni 2019; Meel and Vishwakarma 2020; Asghari et al. 2022).
  - Temporal features: Generally, the propagation of disinformation and true information tend to follow distinct temporal patterns (Yu 2018).

## 5 Disinformation Detection

The literature on detecting disinformation in social networks encompasses a variety of approaches. In this section, we begin by presenting a formal definition of the task of disinformation detection. Then, we explore various approaches found in the literature and categorize them from two distinct perspectives: 1) categorization based on features, and 2) categorization based on algorithms.

### 5.1 Problem Definition

In the literature, various definitions can be found for the problem of disinformation detection (Phan et al. 2023; Shu et al. 2017), characterized by distinctions in the type of disinformation (fake news, rumors, hoaxes, etc) and the datasets used. Some datasets are binary, such as FaKeNewsNet, which labels news articles as fake or real. Others are multiclass, such as LIAR, which assigns six levels of veracity to statements: true, mostly-true, half-true, barely-true, false, and pants-fire. More datasets are discussed in section 6. Moreover, the literature on disinformation detection has explored various methods, but supervised

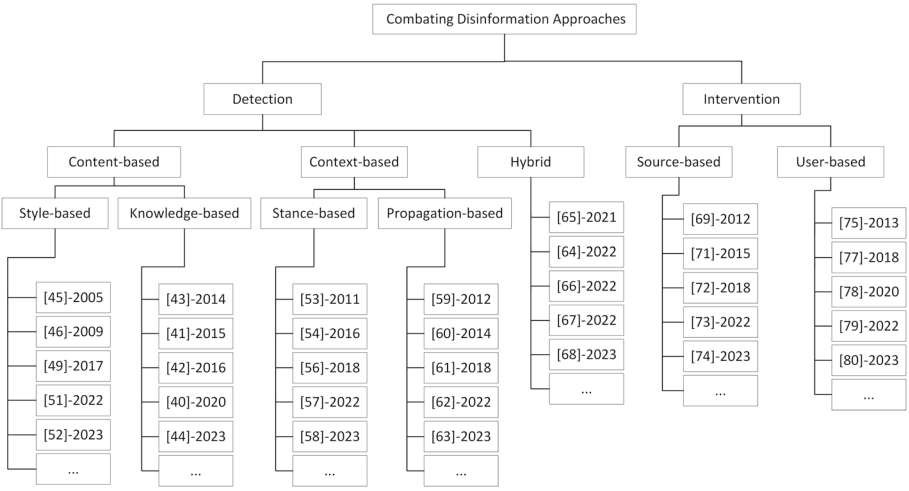


**Fig. 5** A categorization of combating disinformation approaches. In each [i]-j, "i" is the reference number and "j" is the year of the publication

classification approaches have been the most prevalent. Hence, we define the problem of identifying disinformation as follows.

Consider $D = \{d_1, d_2, \ldots, d_N\}$ as a collection of $N$ information items (news, social media posts and more), with each $d_i$ potentially encompassing content-based features (subsection 4.1) and/or context-based features (subsection 4.2). The problem of disinformation detection can be framed as a multi-class classification task, where the goal is to learn a model $F$, that maps each information item $d_i \in D$ to its correct class, using a set of information items whose labels are already known.

## 5.2 Feature-based classification

As mentioned in section 4, various types of features are utilized for disinformation detection. In this section, we propose a categorization, based on the features used by the existing methods. Each method is categorized based on the primary feature it focuses on. In general, there are two approaches to combat disinformation: intervention and detection. The primary objective of detection approaches is to identify and distinguish disinformation from true information. On the other hand, the goal of intervention approaches is to impede the dissemination and spread of disinformation. This classification is depicted in Fig. 5.

### 5.2.1 Detection approaches

Disinformation detection approaches can be categorized into the following three categories:

1. Content-based: As mentioned in subsection 4.1, content-based features are among the most common features used in disinformation detection. Content-based methods are built on the premise that true and false information exhibit distinct textual and visual features. There are two types of content-based methods:

- Knowledge-based methods: Publishers of disinformation often attempt to disseminate false claims through posts on social networks. Therefore, one of the simplest methods involves identifying, verifying, and fact-checking the claims made in these posts. Knowledge-based methods rely on external sources to verify the veracity of claims and assign a truth value to each piece of content (Shu et al. 2017). These methods employ various sources for verification, which can be performed manually, semi-automatically and automatically. Manual methods rely on experts from various fields of expertise to verify claims. In semi-automatic methods, the wisdom of crowd is utilized to detect disinformation. In these methods, individuals can annotate content. Then, by aggregating and analyzing these annotations, an overall assessment of the content veracity is generated. For instance, the Fiskkit[1] website allows users to engage in discussions and annotate the veracity of news articles. Automatic methods conduct the detection process by automatically extracting content-based features and employing machine learning algorithms to classify content. Some of these methods utilize various sources, such as the web (e.g., Wikipedia) and knowledge graphs, to ascertain the accuracy of the content (Shu et al. 2017). Automatic fact-checking using a knowledge graph is a thriving area of research. If successfully implemented, automatic fact-checking using a knowledge graph can serve as a valuable substitute for the laborious and expensive task of manual fact-checking conducted by experts (Shu et al. 2020). Methods in Ciampaglia et al. (2015), Shi and Weninger (2016), You et al. (2014), Wang et al. (2020), Groza (2023) employ a knowledge-based approach to detect disinformation.
- Style-based methods: Similar to knowledge-based methods, style-based methods are also rooted in content analysis. However, while knowledge-based methods focus on evaluating the accuracy of information, style-based methods can also assess the intention behind the dissemination of information. In other words, these methods can determine whether the dissemination of disinformation aims to deceive and mislead the audience or not (Zhou and Zafarani 2020). In disinformation, a distinct writing style is often employed to attract and persuade a wide range of audiences, which differs from the writing style used in true information (Shu et al. 2017). Writing style can be determined by analyzing both textual and visual features. Therefore, style-based methods attempt to differentiate disinformation from true information by extracting and analyzing these features. The writing style can vary across different fields and languages and may also evolve over time (Zhou and Zafarani 2020). Examples of algorithms that utilize writing style to detect disinformation includes (Lesce 1990; Vrij 2005; Mihalcea and Strapparava 2009; Chen et al. 2015; Potthast et al. 2017; Lim 2020; Himdi et al. 2022; Zhou et al. 2023).

2. Context-based methods: In social networks, researchers have access to additional sources beyond content to enhance and reinforce their detection models. Users' comments about content and propagation patterns on social networks often contain valuable and insightful information. Context-based methods leverage the social participation of users and extract diverse information to enhance their detection capabilities. Although context-based methods have received less attention compared to content-based methods, studies demonstrate that incorporating context information can be effective in enhancing the performance of disinformation detection models. In general, these methods can be divided into two categories:

[1] https://fiskkit.com.

- Stance-based methods: In stance-based methods, users' comments are utilized to assess the accuracy of the content. By analyzing users' opinions, it is possible to determine the overall stance towards content. The stance of users can be determined in two ways: explicit and implicit. Explicit stances refer to direct expressions of feelings and opinions, such as "like" and "dislike" reactions. Implicit stances are extracted from comments using natural language processing methods. These methods aim to determine whether the user disagrees, agrees, or remains neutral in relation to a claim or statement (Shu et al. 2017). Users effectively express their sentiments towards content by registering their comments. Studies indicate that disinformation evokes different emotional responses in the audience compared to true information. Therefore, analyzing users' emotions can be effective in the detection of disinformation. Usually, disinformation triggers emotions such as fear, disgust and surprise in users, whereas true information elicits emotions such as sadness, anticipation, happiness and trust (Shu et al. 2020). Hence, users' comments are highly influential, and even without access to the content itself, they can provide valuable insights and general perspectives about the content (Sharma et al. 2019). The methods of Qazvinian et al. (2011), Jin et al. (2016), Mohammad et al. (2017), Qian et al. (2018), Huang et al. (2022), Hamed et al. (2023) analyze the stance of users to detect disinformation.
- Propagation-based methods: Malicious agents aiming to spread disinformation often strive to publish posts and articles rapidly, frequently resorting to bots or fake accounts to expedite the dissemination of content (Shu et al. 2020). The propagation pattern of disinformation differs from that of true information on social networks. For instance, fake political news tends to be disseminated more swiftly and extensively compared to genuine news (Zhou and Zafarani 2020). Hence, analyzing users' profiles and investigating propagation patterns can offer valuable insights to ascertain the authenticity of content (Shu et al. 2020). The methods outlined in Jin et al. (2016), Gupta et al. (2012), Jin et al. (2014), Liu and Wu (2018), Imaduwage et al. (2022), Guo et al. (2023) are primarily rooted in the analysis of information propagation.

3. Hybrid methods: As previously mentioned, content-based methods, particularly style-based methods, are among the most commonly employed approaches for detecting disinformation in social networks, while context-based methods have received relatively less attention. Style-based methods heavily rely on the content of information for their effectiveness. They can be highly effective in the early detection of disinformation and play a crucial role in preventing its further dissemination on social networks. However, if the writing style undergoes significant changes, the efficiency of the model may decrease. In other words, detecting disinformation based on style becomes a cat-and-mouse game, as adversaries can adapt and modify their writing style to evade detection. Every successful detection in the process inspires countermeasures by the disseminators of disinformation (Zhou and Zafarani 2020). On the other hand, propagation-based methods exhibit a more robust performance against malicious entities compared to style-based methods. However, they may be less effective in early detection of disinformation (Zhou and Zafarani 2020). Therefore, hybrid methods leverage a combination of content-based features and context-based features to address the limitations of each method and thereby deliver a more efficient detection model. For instance, the method proposed in Davoudi et al. (2022) analyzes the propagation network and evaluates the stance of users to identify fake news. Other hybrid methods for detecting disinforma-

tion include the approaches presented in Yang et al. (2021), Ran et al. (2022), Liu et al. (2022), Yan et al. (2023).

### 5.2.2 Intervention approach

The goal of the detection approach is to identify disinformation on social networks, which is typically conducted after the news has been written and disseminated. Hence, if the detection process is not carried out promptly, it can result in the widespread dissemination of disinformation, allowing its adverse effects to become more entrenched and lasting. In the intervention approach, the emphasis is on identifying the factors that contribute to the spread of disinformation. Intervention approaches aim to identify and address disinformation before it spreads widely. These methods can be broadly categorized into the following two categories based on the identification of the type of agent involved:

1. Source-based methods: The first category focuses on the sources of information, including the authors and publishers, as well as individuals who are quoted or referenced in the content. In general, evaluating the credibility and validity of sources serves as an indirect method for identifying disinformation. In other words, information obtained from unreliable news sources is often considered as potentially false, even though there may be cases where the information is true (Zhou and Zafarani 2020). In general, source-based methods aim to mitigate the spread of disinformation from unreliable sources by evaluating the credibility and trustworthiness of information sources. This approach can indeed play a significant role in preventing the dissemination of disinformation by targeting the sources. Methods presented in Shah and Zaman (2012), Karamchandani and Franceschetti (2013), Xu and Chen (2015), Baly et al. (2018), Fraszczak (2022), Zhang et al. (2023) focus on identifying and evaluating the sources of information to determine their credibility and reliability.
2. User-based methods: On social media platforms, information is disseminated by users. Intuitively, users with lower credibility are more likely to disseminate disinformation compared to users with higher credibility. These users are more susceptible to being exposed to such information. In general, users with low credibility can be considered as potentially malicious users, while users who are vulnerable to the spread of false information can be seen as regular or normal users (Zhou and Zafarani 2020). Therefore, identifying and deactivating accounts of malicious users can be helpful in preventing the dissemination of disinformation. The methods presented in Di Pietro et al. (2013), Cresci et al. (2014), Chen and Wu (2018), Jabardi and Hadi (2020), Li et al. (2022), Mughaid et al. (2023) aim to prevent the spread of false information by identifying malicious users.

### 5.2.3 Discussion

In this section, we explored the strategies for addressing disinformation, and categorized them into two groups: detection methods and intervention methods. As previously mentioned, detection methods try to identify and stop disinformation that has already spread. These methods prevent its further dissemination. In contrast, intervention methods aim to preempt the creation and propagation of disinformation by identifying and eliminating the contributing agents, such as fake users (Di Pietro et al. 2013; Mughaid et al. 2023) or unreliable sources (Karamchandani and Franceschetti 2013; Zhang et al. 2023). Naturally,

intervention methods, which operate by eliminating contributory agents, have the potential to outperform detection methods by preempting the production and propagation of disinformation. However, in intricate settings such as social networks, replete with diverse forms of interactions, devising these methods can be notably challenging. This is especially true as the accurate and precise execution of these intervention methods bears significant importance in identifying and thwarting the influence of agents like fake users. Any missteps in their execution could result in substantial user dissatisfaction. Because of these reasons, detection methods currently enjoy greater popularity.

We scrutinized these methods, categorizing them into three distinct groups based on the type of features used for detection: content-based, context-based, and hybrid. As mentioned above, initial endeavors revolve around content-based methods. While content, particularly textual content, constitutes a crucial and fundamental element of information, it may not suffice for the accurate identification of disinformation (Bondielli and Marcelloni 2019). This is due to the variability of content-based features, such as linguistic cues and writing style, across different domains, languages, and their evolvement over distinct time periods (Zhou and Zafarani 2020). Furthermore, concomitant with the advancements in the precision of content-based models, malicious actors can adapt and modify content attributes, such as writing style, by studying the operation of these models. Consequently, they may circumvent these models successfully (Zhou and Zafarani 2020). Hence, the advancement of context-based models, which remain unaffected by content-based features and resilient against manipulative factors, offers a promising solution to these challenges. In practice, spreading disinformation has limited influence over users' positions regarding the shared content and the mechanisms of information propagation. However, it's essential to underscore that context-based methods are not suitable for early detection (Zhou and Zafarani 2020). To capture context features such as propagation graphs or user stances, a certain amount of time and the accumulation of substantial background information are required. Consequently, a pragmatic approach lies in amalgamating content-based and context-based methods (Yang et al. 2021; Yan et al. 2023). This dual approach not only enhances the detection model's resilience against malicious attacks but also maintains a commendable performance in early detection scenarios.

### 5.3 Algorithm-based categorization

In this section, we present a categorization of disinformation detection methods, based on the algorithms employed. Figure 6 illustrates this categorization.

Initially, the methods are classified into two categories based on whether they utilize machine learning techniques or not. The first category comprises methods that do not employ machine learning algorithms, such as the rhetorical approach, which involves manipulating sentence components to enhance the persuasive impact of speech. Rhetorical Structure Theory (RST) is a theoretical framework for organizing text that explains the relationships that exist between different parts of a discourse. Some researchers have employed RST in their efforts to detect fake news. RST analyzes the semantic role of textual units and subsequently identifies the main idea or central theme of the text. Subsequently, the identification of fake news is based on the evaluation of text coherence and structure (Oshikawa et al. 2018). Cognitive psychological analysis of the text by experts is another method employed to assess the accuracy of information. The methods presented in Kumar and Geethakumari (2014), Rubin et al. (2015), Shu et al. (2017), Vedova et al. (2018) utilize non-machine-learning approaches for disinformation detection.

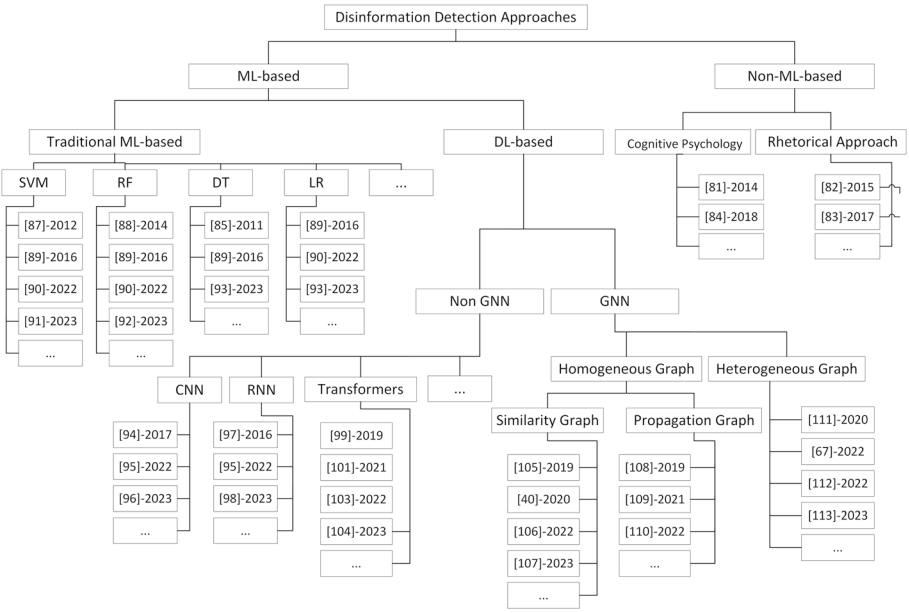


**Fig. 6** Algorithm-based categorization of disinformation detection approaches. In each [i]-j, "i" is the reference number and "j" is the year of publication

The second category encompasses methods that utilize machine learning (ML) algorithms. Initial ML-based approaches primarily concentrate on determining credibility in social networks (Castillo et al. 2011) and detecting fraud in computer communication (Zhang et al. 2012). These approaches have shown promising results. Consequently, machine learning methods are widely employed in numerous studies for disinformation detection (Bondielli and Marcelloni 2019). These methods can be categorized into two distinct groups:

1. The first category consists of methods that utilize traditional machine learning algorithms, such as support vector machine (SVM) (Afroz et al. 2012; Briscoe et al. 2014; Giasemidis et al. 2016; Elyassami et al. 2022; Kini 2022), random forest (RF) (Briscoe et al. 2014; Giasemidis et al. 2016; Elyassami et al. 2022; Sharma et al. 2023), decision tree (DT) (Castillo et al. 2011; Giasemidis et al. 2016; Kishwar and Zafar 2023) and logistic regression (LR) (Giasemidis et al. 2016; Elyassami et al. 2022; Kishwar and Zafar 2023).
2. The second category comprises methods that leverage deep learning (DL) models. In recent years, deep learning models have experienced unparalleled popularity due to their exceptional performance in various tasks, including text mining and natural language processing. These models possess a significant advantage over traditional machine learning methods. Indeed, in traditional machine learning models, features are extracted manually. It can be tedious, time-consuming, and potentially result in biased features. In deep learning models, feature extraction is done automatically, eliminating the need for manual feature engineering. Deep learning models have the ability to learn hidden features that may not be readily apparent to human experts (Bondielli and Marcelloni

2019). This paper primarily focuses on methods that employ graph neural networks (GNNs). So, we divide deep learning models into two categories:

(a) The first category includes DL models that do not utilize graph neural networks. Convolution neural networks (CNN) (Chen et al. 2017; Kadek et al. 2022; Vishwakarma et al. 2023), recurrent neural networks (RNN) (Kadek et al. 2022; Ma et al. 2016; Kishore and Kumar 2023) and transformers models (Slovikovskaya 2019; Qazi et al. 2020; Hande et al. 2021; Schütz et al. 2021; Rai et al. 2022; Praseed et al. 2023) are among the most commonly used deep learning models for detecting disinformation.

(b) The second category comprises DL models that utilize GNNs. As highlighted in subsection 3.2, GNNs have demonstrated remarkable success in a wide range of problems, including those that are not inherently graph-related. The success of GNNs has motivated researchers to place particular emphasis on these models for identifying disinformation in social networks. Based on the type of graph constructed, GNN-based methods can be divided into two categories:

- Homogeneous graph-based methods: In these methods, a homogeneous graph is constructed. Two different approaches are commonly employed to construct this graph:
  - The first approach involves creating a similarity graph by utilizing a similarity measure between the content of posts, user comments, or user features. In the methods proposed by Wang et al. (2020), Benamira et al. (2019), Weizhi et al. (2022), Cui et al. (2023), a similarity graph is created.
  - The second approach involves creating a propagation graph for each news event by considering the source post and its subsequent reposts. In the methods of Huang et al. (2019), Song et al. (2021), Song et al. (2022) a propagation graph is created.
- Heterogeneous graph-based methods: Social networks are characterized by heterogeneous graphs, encompassing various types of nodes (users, posts, comments, etc) and edges (followers, friends, reposts, etc). So, in certain methods, a heterogeneous graph is constructed and processed, where each post node is subsequently classified as either real or fake. The methods presented in Liu et al. (2022), Ren et al. (2020), Zheng et al. (2022), Azarijoo et al. (2023) serve as examples of heterogeneous graph-based approaches.

#### 5.3.1 Discussion

In this section, we categorized disinformation detection methods based on the type of algorithms they employ. The substantial strides made in artificial intelligence in recent years have elevated machine learning algorithms to a prominent position across various domains, and the realm of disinformation detection is no exception. This is particularly pertinent given the escalating significance of combatting the dissemination of disinformation, which has emerged as a substantial challenge in contemporary human societies. Traditional machine learning algorithms, although capable of delivering satisfactory performance, entail manual feature extraction, which is a labor-intensive and error-prone process, demanding substantial time and resources. Conversely, deep learning algorithms possess the capacity to automatically extract intricate features, rendering them superior to traditional ML techniques (Bondielli

and Marcelloni 2019). This class of methods encompasses various types of deep neural networks designed to extract valuable features and proficiently identify disinformation through diverse strategies. One of the pivotal determinants influencing the efficacy of deep models is the approach to data modeling, as inaccuracies in data modeling can detrimentally impact the model's performance. Typically, non-GNN models excel in Euclidean spaces, but their performance diminishes in non-Euclidean spaces like graphs, as they struggle to capture intricate relationships. In contrast, graph-based modeling proves to be a potent and effective technique for modeling complex data structures. When it comes to analyzing non-Euclidean data, GNNs emerge as the preeminent tools in the arsenal (Chehreghani 2022).

The problem under investigation, disinformation detection, can be approached as a graph-based problem in several facets: the platforms through which this information proliferates (such as social networks and the web) are essentially complex networks, ideally modeled as graphs, some of the features employed in detection methods, like propagation graphs, inherently have graph structures, while even non-graphical features such as text content (comprising words and their relationships) can be transformed into graph-like structures. Hence, it appears that GNNs offer a superior choice for the development of disinformation detection models. By amalgamating initial features with structural information, they stand as a potent option to attain commendable performance.

The effectiveness of GNNs is contingent on the manner in which graphs are constructed from the available data. Existing models employ diverse approaches to generate graphs. For instance, the propagation graph is created based on the propagation of information, with the model having minimal direct involvement in graph generation. Conversely, the similarity graph encompasses various approaches, including criteria related to the similarity between information content or the features of users' profiles. Diverse models leverage different criteria for crafting similarity graphs, yet, in general, one of the prevailing techniques for generating similarity graphs involves utilizing the k-nearest neighbors (KNN) method rooted in content similarity (Benamira et al. 2019). In aforementioned approaches (propagation graph and similarity graph), the initial feature vector of nodes may encompass textual or visual content representations, along with specific features extracted from users' profiles. These graphs are homogeneous in nature, as they consist of nodes of the same type, whether they represent content or users. Sometimes, taking into account both the published content, the publishing users, and the diverse interactions between them can furnish the model with richer and more comprehensive information (Liu et al. 2022). Moreover, social networks inherently constitute a heterogeneous graph, incorporating distinct types of nodes and edges. Consequently, there has been a growing trend towards exploring heterogeneous graph-based models in recent years. Given the proficient performance of GNNs and their intrinsic capacity to frame the disinformation detection problem as a graph-based challenge, they appear to be highly suitable for disinformation detection. However, it's imperative to remember that the precise and effective modeling of information as a graph can profoundly impact the performance of these models. Incorrect modeling can result in a significant deterioration of their performance.

### 5.3.2 GNN-based disinformation detection methods

In this section, we examine and compare several recent GNN-based approaches for disinformation detection, categorizing them into three distinct categories: similarity graph-based methods, propagation graph-based methods, and heterogeneous graph-based methods.

*Similarity graph-based methods*

Benamira et al. (2019) proposed a straightforward content-based semi-supervised method that employs GCNs and GATs to detect fake news. The method consists of three steps: First, the GloVe method (Pennington et al. 2014) is utilized to extract the feature vector of news articles. Next, the $k$-nearest-neighbor method (where $k$ is set to 4) is employed to establish a graph among articles based on word embedding similarities. Finally, the GCN or GAT models are leveraged to classify the nodes as either fake or real.

Wang et al. (2020) proposed a knowledge-driven multimodal graph convolutional network (KMGCN) to effectively capture the background knowledge embedded within the text content of a post, by jointly modeling textual, visual and conceptual information. Rather than treating text as a linear sequence of words, this model represents the content of each post as a graph to model non-consecutive phrases and enhance the capture of semantic composition. Furthermore, object detection techniques are employed to extract visual objects as visual words, and a knowledge graph is utilized to extract complementary semantic information in the form of knowledge concepts. To construct the graph, the words, visual words, and knowledge concepts are considered as the graph nodes, and weighted edges are established using pointwise mutual information (PMI). A 2-layer GCN and global pooling are subsequently applied to aggregate the nodes within each graph, resulting in an embedding vector for each post. Finally, the embedding vector is fed into a binary classifier, which calculates the predicted probability of whether the news is fake or not. Overall, this model offers an effective approach to integrate diverse forms of information, in order to enhance the detection of fake news.

One of the main challenges in disinformation detection is that many models are specifically designed to operate within a particular domain. In other words, these models are trained on datasets specific to certain domains (e.g., sports or politics), which limits their ability to effectively learn domain-specific features. Consequently, they may not perform well when applied to data from other domains. To tackle this challenge, Yuan et al. (2021) proposed a novel domain-adversarial and graph-attention neural network (DAGA-NN) model, which has the capability to learn domain-invariant features of fake news across different events and domains. The model simultaneously preserves the relationships between real and fake samples within the same domain, ensuring high performance in the identification process. To construct the graph, each news item is considered as a node, and an edge is created between corresponding nodes if the cosine similarity between their feature representations exceeds a predefined threshold value. DAGA-NN employs a multimodal feature extractor to obtain feature representations. To detect fake news across events/domains, it introduces a domain discriminator based on the concept of domain-adversarial networks. This discriminator engages in a minimax game with the feature extractor, aiming to learn domain-invariant features. DAGA-NN further utilizes graph attention networks (GAT) followed by a fully connected layer to classify news items as either real or fake. Overall, this model offers a promising solution to the challenge of detecting disinformation across multiple domains.

When attempting to detect disinformation in short news articles, relying solely on their content and linguistic features may not produce satisfactory performance. To address this challenge, a multi-depth graph convolutional networks framework (M-GCN) has been proposed (Hu et al. 2019) that leverages user profiles to enhance the analysis of news articles. In the M-GCN model, each news text is treated as a node within the graph structure. Additionally, a speaker's profile, including attributes such as their political party, the news topic, or their home state, is considered. If two nodes share the same value for a given profile attribute, they are connected via an edge. This

approach extends the capabilities of GCNs by incorporating multi-scale information about the neighbors of each node. To obtain the final representation for detecting fake news, the node features and the outputs of the multi-depth GCN blocks are combined using an attention mechanism. This enables the model to proficiently capture pertinent information from both the news text and the corresponding user profiles, thereby enhancing its capacity to detect disinformation in concise news articles.

FauxWard (Shang et al. 2020) is a specifically designed model for the identification of fauxtography on social media. This model addresses situations where the combination of both images and text is employed to convey false or misleading information. The challenge of detecting fauxtography necessitates not only assessing the accuracy of the image and text components of a post separately, but also scrutinizing the accuracy of the correlation between the image and the text. Unlike other methods, FauxWard is considered content-free, implying that it does not analyze the visual or textual contents of the post. Instead, it employs the GCN model to explicitly investigate the intricate information derived from a user comment network that is linked to a social media post. To accomplish this, FauxWard initiates the construction of the user comment graph by utilizing the reply relationship among the comments associated with a post. Subsequently, it proceeds to extract the intricate information from each user comment node, by considering a set of linguistic and semantic attributes. Finally, the GCN model is used to process the constructed graph and classify posts as either fauxtography or not-fauxtography. This model demonstrates robustness against sophisticated content creators who possess expertise in crafting and disseminating misleading fauxtography content across social media platforms.

The IIJIPN (intra-graph and inter-graph joint information propagation network) model (Cui et al. 2023) utilizes diverse relationships between words to disseminate information within and among graphs. To start, sequential, syntactic, and semantic features are extracted from the news text and utilized to construct three types of graphs: sequential, syntactic, and semantic graphs ($G_{seq}$, $G_{syn}$, and $G_{sem}$). In these graphs, words are considered as nodes, and prior to creating the sequential and syntactic graphs, the words are normalized and then processed using PMI and CoreNLP. Next, semantic features are extracted by capturing the correlations between words that are related to the topic. Subsequently, intra-graph information propagation is performed independently for each graph using the gated graph neural network (GGNN) to update the node embedding vector. The composed graph $T = (G_{seq}, G_{syn}, G_{sem})$ is then transformed into $T_{intra} = (G'_{seq}, G'_{syn}, G'_{sem})$ after intra-graph information propagation. Inter-graph information propagation is then conducted to aggregate information between different graphs. After the intra-graph information propagation is completed for each graph, any two updated graphs are selected to enable inter-graph word interaction. The graph $T_{intra} = (G'_{seq}, G'_{syn}, G'_{sem})$ is then transformed into $T_{inter} = (G''_{seq}, G''_{syn}, G''_{sem})$. The graph attention mechanism is then employed to calculate the final graph representation $G_f$, which is sent to the node attention mechanism to obtain the node representation $w_f$. Finally, the obtained node representation is fed into a classifier to determine whether the news is fake or real.

Weizhi et al. (2022) focused on evidence-based fake news detection, where multiple pieces of evidence are utilized to examine the truthfulness of a claim. They proposed a unified graph-based semantic structure mining framework (GET). Unlike other approaches that consider claims and evidences as sequences, they represented them as graph-structured data and utilized neighborhood propagation to capture long-distance semantic dependencies among scattered relevant snippets. GET can be primarily divided into four modules:

1. Graph construction: In order to capture the long-distance dependency of relevant information, the authors adopted a sliding window approach to convert the original claims and evidences into graphs. Words are treated as nodes in the graph representation. Next, the center words in each window were connected to the other words within the window. This step establishes connections between the center word and the surrounding words, capturing the local context and relationships within the neighborhood of the center word. Furthermore, to capture the long-distance dependency, they merged all occurrences of the same words into a single node in the graph. This consolidation explicitly gathers the local contexts of these word instances: if there is one claim and $n$ evidences, the framework constructs $(n + 1)$ graphs.
2. Graph-based semantics encoder: To extract the long-distance semantic dependency, they proposed employing GNNs as the semantic encoder. They utilized gated graph neural networks (GGNN) to facilitate neighborhood propagation on both claim and evidence graphs, allowing nodes to effectively capture contextual information.
3. Semantic structure refinement: As evidence often contains redundant information that can potentially divert the model's attention towards unimportant features, it is advantageous to identify and filter out this redundancy, thereby, obtain refined semantic structures. To achieve this objective, the authors proposed to compute a redundancy score for each node. The redundancy arises not only from the self-information contained within each node but also from the contextual information present in the neighborhood of the graphs. For instance, if a claim can be verified by a specific snippet within an evidence, the remaining segments (including the context of the snippet) will be considered redundant. The authors employed a 1-layer GGNN to calculate the redundancy scores, considering both self-information and contextual information in the computation process. Additionally, they conducted semantic structure refinement only on evidences, as claims typically tend to be shorter in length. Lastly, they stacked the semantic structure refinement layer on top of a semantics encoder, creating a unified module known as evidence semantics miner (ESM). This module effectively captures long-distance semantic dependencies and reduces redundant information.
4. Attentive graph readout layer: The authors integrated all node embeddings (word embeddings) into general graph embeddings (claim and evidence embeddings). Then they extended claim and evidence representations by concatenating them with corresponding information vectors. Finally, they combined the claim and evidence embeddings into a unified representation by concatenating them, and then passed this representation through a multi-layer perceptron to generate the veracity prediction label.

*Propagation graph-based methods*

Autef et al. (2020) proposed a graph classification approach using a simple GNN to detect fake news on Twitter. The method involves constructing a propagation tree for each news item on Twitter, utilizing both text features (extracted using the BERT transformer-based language model) and user features. Subsequently, the propagation tree is classified separately using three different models: GCN, GAT and GraphSage. Lin et al. (2020) proposed a model for detecting rumors on social networks that incorporates textual, propagation, and structure information. The model consists of three components: encoder, decoder, and detector. The encoder utilizes a GCN to update the representation of the initial text through the propagation graph, allowing it to learn both textual and propagation features. The encoded representation is then fed into the decoder, which utilizes an AutoEncoder to learn overall structure information. Simultaneously, the

detector employs the encoder's output to classify events as either fake or not. These three modules are jointly trained to enhance the model's effectiveness.

Huang et al. (2019) proposed a hybrid model for detecting rumors on social media by considering three aspects: users, content and propagation graph. The model consists of three modules: user encoder, propagation tree encoder and integrator. The user encoder utilizes GCN to capture and model users' attributes and behaviors. These attributes and behaviors are represented as a co-occurrence relationship graph. In the co-occurrence graph, the nodes represent users. The propagation tree encoder employs recurrent neural networks (RNN) to encode the structure of the rumor propagation tree into a vector representation. Finally, the integrator combines the output of two previous modules using a fully connected layer to classify rumors.

TGNF Song et al. (2021) is a GNN-based model that tackles the limitation of static network views, which assume that the complete information propagation network structure is available prior to executing learning algorithms. To model temporal propagation patterns of news, it leverages temporal graph attention networks (TGAT) Xu et al. (2020) to capture dynamic structure, content semantics, and temporal information during the news propagation process. It models news propagation using continuous-time dynamic graphs (CTDG). Additionally, TGNF incorporates the temporal difference network (TDN) inspired by adversarial learning. This enables the model to capture the variations in information between interactions, rather than emphasizing similar ones. Propagation graphs of news are constructed at different time steps and processed to determine whether their label is fake or real.

The Bi-GCN model Bian et al. (2020) combines two GCN networks to process the rumor propagation tree in two different directions: top-down and bottom-up. First, a propagation tree is constructed for each news event. The model utilizes a GCN with a top-down directed graph of rumor diffusion to learn the patterns of rumor propagation. Additionally, it utilizes another GCN with an opposite directed graph of rumor diffusion to capture the structures of rumor dispersion. Moreover, the source post information is integrated into each layer of GCN to enhance the influence from the roots of rumors. Finally, the representations from these two networks are combined using a fully connected layer, and the final label is determined.

Uncertainty poses another challenge in disinformation detection. Uncertainty in the propagation graph of disinformation is inevitable. On one hand, disinformation producers often manipulate others by generating fake supporting tweets or eliminating opposing voices in order to evade detection. On the other hand, the available graph may only represent a portion of the actual propagation structure and may include noisy relations, leading to inherent uncertainty Wei et al. (2021). To address this challenge, Wei et al. (2021) proposed a novel approach called edge-enhanced Bayesian graph convolutional network (EBGCN) for rumor detection. This model addresses the uncertainty issue in the propagation structure by modeling it from a probability perspective. Similar to BiGCN (Bian et al. 2020), for each news event, two weighted propagation graphs (top-down and bottom-up) are constructed. The feature vector of each node is then extracted using the TF-IDF method. These two propagation graphs are processed separately by two GCNs. The embedding vectors obtained from the GCNs are then aggregated to classify the news. The core idea of EBGCN is to dynamically control the message-passing process based on the prior belief of the observed graph. This approach replaces the fixed edge weights in the propagation graph with adaptive weights. In each iteration, edge weights are inferred by the posterior distribution of latent relations according to the prior belief of node features in the observed graph. Then, graph convolutional layers

are used to aggregate node features by considering various adjacent information on the refining edges.

Jiang et al. (2021) identified two problems in detecting false information:

- Using two uni-directional networks separately (top-down and bottom-up) restricts the interaction between nodes.
- There is a gap between training and target representations, as most approaches only learn node-level embeddings and rely on average or max pooling to obtain whole graph embeddings.

To address these issues, they proposed the L-GAT method. L-GAT views all edges as bi-directional and employs a unified graph attention network to aggregate information, disregarding the distinction between top-down and bottom-up information flow. Additionally, two landscape nodes are added to the original graph to capture global information. These nodes' embeddings are trained using an end-to-end method to bridge the gap between the training and target representations. L-GAT utilizes the GAT network as the backbone model with three modifications: the inclusion of virtual global nodes (landscape nodes), root feature enhancement, and data augmentation. First, L-GAT incorporates two virtual landscape nodes, namely the virtual extension node (VEN) and the virtual RBF-kernel node (VRN). These nodes are introduced to facilitate the learning of global node embeddings in an end-to-end manner. Second, the root node holds significant importance in rumor detection. Therefore, L-GAT employs root feature enhancement, where for each layer of L-GAT, all node embeddings are concatenated with the representation of the previous layer's root node. This allows the model to capture and incorporate the crucial information from the root node in each layer. Finally, to mitigate the risk of overfitting, L-GAT employs the DropTree procedure. In the DropTree method, a node and its subtrees are removed, ensuring that the remaining trees maintain connectivity and are still rooted by the original root node. This approach helps in regularizing the model and preventing overfitting.

Zhiyuan et al. (2020) introduced two rumor detection methods based on the attention mechanism in graph neural networks. They first constructed the propagation graph based on the non-sequential propagation structure of posts on social networks. If one tweet is a response to another, two directed edges are created between them. They then introduced a gated graph neural network (GGNN), called PGNN, to learn node embeddings in the propagation graph. PGNN iteratively updates node embeddings by exchanging information between neighboring nodes, via relation paths within a limited number of time steps. The authors proposed two different models for the rumor detection task: GLO-PGNN and ENS-PGNN. GLO-PGNN identifies rumors based on the global embedding of the propagation graph. It uses a fully connected layer to classify the global embedding. ENS-PGNN employs an ensemble approach. It calculates the prediction probability for each node in the propagation graph and then aggregates these probabilities to obtain the final result.

Malhotra et al. (2020) proposed a method for rumor detection that constructs the propagation graph at the user level, where each user involved in spreading news is considered as a node and their profile features are utilized as node features. The authors extracted 12 profile features from Twitter, including the number of followers, the number of friends and the number of tweets. They utilized a 2-layer GCN to learn the embedding of the propagation graph. In addition, they processed the text of the source tweets using RoBERTa-based word embedding Liu et al. (2019), to obtain the feature vector of the source tweet. Lastly,

they fed the acquired graph embedding vector and the feature vector of the source tweet into a fully connected layer for classification purposes.

Thota et al. (2023) proposed a propagation-based method for early detection of rumors by capturing influential features and learning dynamic structures of their propagation. Their model consists of four modules: feature extraction, pattern matching, structure reconstruction, and rumor detection. Within the propagation graph, individual users are depicted as nodes, and node representations are generated using a 2-layer GCN. The edges within the graph are fed into the pattern matching algorithm, which recursively analyzes the data stream and generates a list of matched patterns along with their corresponding occurrence times. Each edge consists of two nodes and the timestamp indicates the time of its generation. The two most frequently occurring patterns in the stream are the "star" and "path" structures. The "star" structure branches out into multiple paths, while the "path" structure follows a single propagation path within a subgraph. The model identifies a group of patterns by considering both behavioral and structural features for stance classification over subgraphs. Lastly, the sequence of graph snapshots over time is recursively integrated to reconstruct the structure, enabling the differentiation between rumors and non-rumors.

Lu and Li (2020) proposed a model called graph-aware co-attention network (GCAN) for predicting fake news based on the source tweet and its propagation through users. GCAN consists of five components. The first component extracts user characteristics that quantify the extent of a user's engagement in online social networking. User features are derived from their profiles and interactions within the social network. The second component encodes the news story by generating word embeddings for the source tweet. The third component models and represents how the source tweet propagates through users by utilizing their extracted characteristics. It employs convolutional and recurrent neural networks to learn the embedding of the retweet propagation, based on user features. Additionally, a fully-connected graph is constructed to model the potential interactions among users, and the graph convolution network is employed to learn the graph-aware representation of user interactions. The fourth component employs dual co-attention mechanisms to capture the correlation between the source tweet and retweet propagation, as well as the co-influence between the source tweet and user interaction. The last component generates the detection outcome by concatenating all the learned embeddings.

*Heterogeneous graph-based methods*

Ren et al. (2020) proposed a model known as AA-HGNN (adversarial active learning-based heterogeneous graph neural network), which utilizes a hierarchical attention mechanism to learn node representations in a heterogeneous information network. Furthermore, it incorporates an active learning framework to improve the performance of learning, particularly in situations where labeled data is limited. The heterogeneous graph consists of three types of nodes: creators, news articles, and subjects. The edge-set includes "write" links connecting creators and news articles, as well as "belongs to" links connecting news articles and subjects. The AA-HGNN model utilizes a hierarchical graph attention neural network (HGAT) as its core module. HGAT comprises a two-level attention mechanism for both node-level attention and schema-level attention. The node-level attention mechanism learns the weights of neighbors belonging to the same type and aggregates them to derive the type-specific neighbor representation. The schema-level attention allows the HGAT model to capture the information regarding node types and achieve an optimally weighted combination of type-specific neighbor representations. HGAT is employed in two major components: HGAT-based classifier and HGAT-based selector. The HGAT-based classifier utilizes both labeled and unlabeled data to predict labels for unlabeled news article nodes. The HGAT-based selector assesses the quality of predicted labels and chooses high-value

candidates using a query strategy. The two components of the model collaborate in an adversarial and active manner to enhance the quality of predicted labels and select superior candidates, thereby improving learning performance. The node representations obtained from the HGAT encompass both structural and node content information of news article nodes.

Huang et al. (2020) proposed a method for rumor detection, that relies on a tweet-word-user heterogeneous graph, constructed using text contents and source tweet propagations. The graph comprises nodes representing words, tweets, and users. The edges are established based on word co-occurrence, word and tweet frequency, as well as user behavior such as retweeting or replying to tweets associated with the source tweet. In order to capture both the global semantic relations of text contents and the information related to source tweet propagations, the graph is decomposed into tweet-word and tweet-user subgraphs. These subgraphs are subsequently combined using a subgraph-level attention mechanism. Node representations are learned using an attention mechanism and then fed into a fully connected network for classification.

Factual news graph (FANG) (Nguyen et al. 2020) is a machine learning approach that operates on an undirected heterogeneous graph network. This graph includes three different types of nodes: news articles, news sources, and social media users. There are four different types of edges in the graph: followership (between users), citation (between news sources), publication (between news sources and news articles) and stance (between users and news articles). The publication and stance edges in the graph are time-dependent, while the other edges are independent of time. To capture the representations of these different entities within the graph, FANG leverages GraphSage. By employing GraphSage, FANG generates general embeddings for each type of nodes, which can be effectively utilized for various tasks. To accomplish this, FANG uses three different cost functions.

- The first cost function, known as the unsupervised proximity loss, is founded on the principle that nodes that are connected in the graph are likely to exhibit similar behavior. Consequently, this implies that news sources that publish similar articles and users who hold similar opinions on articles should be represented in comparable manners.
- The second cost function, referred to as the self-supervised stance loss, shares similarities with the first one but specifically emphasizes the connections between users and news articles. If a user holds a particular opinion about an article, his/her embeddings should exhibit similarity to the embeddings of other users who share the same opinion.
- The third cost function, named as the supervised fake news loss, is used to directly optimize the task of detecting fake news. To achieve this, FANG generates an embedding for each news article by combining both the article's own embedding and the embedding of the news source that has published it.

FANG can be trained rapidly and enables predictions to be made on new data without requiring the entire graph to be re-processed.

Rumor detection methods commonly utilize text semantics and propagation structure to automatically classify rumors. However, they often overlook the impact of false or irrelevant interactions within the propagation structure, leading to reduced accuracy. Moreover, most existing methods fail to effectively extract crucial clues from user comments on social networks. To tackle these challenges, Liu et al. (2022) introduced a social network rumor detection method, called DAGCN (dual attention graph convolutional network). DAGCN combines a dual attention mechanism with a graph convolutional network to enhance the detection of rumors in social networks. It constructs a heterogeneous propagation graph for

each event by utilizing the comment-retweet relationship. In this graph, different events are represented as nodes, while the edges represent the connections between users. The model comprises three primary modules:

1. The first module is the interactive text semantic feature module. It employs the multi-head self-attention mechanism to calculate and merge the source tweets or microblogs with comments and repost content. This process generates a new feature to represent each source post or event.
2. The second module is the anti-interference propagation structure feature module. It utilizes the propagation graph to represent various data elements such as source posts, source microblogs, forwarded microblogs, comments and users. Structural information is then extracted using GCN. The attention mechanism is employed to decrease the weight of inaccurate edges within the propagation graph. This reduction in weight helps to mitigate interference and enhance the accuracy of rumor detection.
3. The third module is the rumor prediction classification module. It combines the interactive text semantic features and anti-interference propagation structure features to generate a novel representation for each event. Finally, the model employs a fully connected neural network to predict the final labels for the rumor detection task.

Zheng et al. (2022) proposed an approach, called multi-modal feature-enhanced attention network (MFAN), for multimedia rumor detection. This approach incorporates textual, visual and social graph features, enabling a joint modeling of multiple modalities for improved performance. To represent user behaviors on social media, the authors established a heterogeneous graph $G = \{V, A, E\}$, where $V$ includes user nodes, comment nodes and post nodes, $A$ is an adjacency matrix between nodes to describe their relationships, and $E$ is the set of edges. For extracting semantic features from sentences in each post, they utilized a convolutional neural network. Additionally, to extract image features, they employed the ResNet50 (He et al. 2016) architecture. To address the missing links issue, they inferred hidden links between nodes by considering feature similarity. They proposed to enhance both the graph topology and aggregation procedure based on the signed attention-based GAT to extract better social graph features. Subsequently, they employed the co-attention mechanism to capture the mutual information across different modalities. This mechanism allowed them to learn attention weights between various modal features, enhancing the cross-modal features by emphasizing their relevance and interactions. Finally, they concatenated the enhanced multi-modal features obtained from the previous steps to create a fused representation. This fused representation was then used for classification tasks.

A summery of key properties of the studied methods are presented and compared in Tables 1 and 2. This comparison reveals that:

- GNN is a novel technique for disinformation detection, with its first research being presented in 2019. The popularity of this technique is on the rise due to its impressive performance.
- GCN and GAT stand out as the most widely utilized graph neural networks.
- The majority of methods employ the propagation graph, which is a homogeneous graph, placing them in the category of context-based methods.
- The majority of methods rely on textual features, and user profile features are also commonly utilized in many approaches. However, features such as comments, semantic characteristics and temporal aspects have been given less consideration.

**Table 1** GNN-based disinformation detection methods (Part 1). ACC stands for accuracy

| Ref. | Year | GNN Type | Graph | Features | Approach | Setting | Disinformation Type | Dataset | Performance |
|---|---|---|---|---|---|---|---|---|---|
| Benamira et al. (2019) | 2019 | GCN GAT | Similarity | Textual | Content-based | Semi-supervised | Fake news | Custom DS | ACC: 0.849 |
| Hu et al. (2019) | 2019 | GCN | Similarity | Textual, Profile | Content-based | Semi- supervised | Fake news | LIAR | ACC: 0.492 |
| Autef et al. (2020) | 2019 | GCN GAT GraphSage | Propagation | Textual, Profile | Context-based | Supervised | Fake news | Twitter15<br>Twitter16 | ACC: 0.690<br>ACC: 0.750 |
| Huang et al. (2019) | 2019 | GCN | Propagation | Textual, Profile | Context-based | Supervised | rumor | Twitter15<br>Twitter16 | ACC: 0.752<br>ACC: 0.773 |
| Han et al. (2020) | 2020 | GraphSage | Propagation | Profile, Temporal | Context-based | Supervised | Fake news | PolitiFact<br>GossipCop | ACC: 0.803<br>ACC: 0.825 |
| Bian et al. (2020) | 2020 | GCN | Propagation | Textual, Profile | Context-based | Supervised | rumor | Weibo<br>Twitter15<br>Twitter16 | ACC: 0.961<br>ACC: 0.886<br>ACC: 0.880 |
| Wang et al. (2020) | 2020 | GCN | Similarity | Textual, Visual, Semantic | Content-based | Supervised | Fake news | Weibo<br>PHEME | ACC: 0.886<br>ACC: 0.876 |
| Shang et al. (2020) | 2020 | GCN | Similarity | Comments | Context-based | Supervised | faux | Reddit<br>Twitter | ACC: 0.754<br>ACC: 0.711 |
| Zhiyuan et al. (2020) | 2020 | GAT | Propagation | Textual | Context-based | Supervised | rumor | PHEME | Macro-F1: 0.753 |
| Bai et al. (2021) | 2020 | GCN | Propagation | Textual | Context-based | Supervised | rumor | PHEME | ACC: 0.841 |
| Malhotra et al. (2020) | 2020 | GCN | Propagation | Textual, Profile | Context-based | Supervised | rumor | Twitter15<br>Twitter16 | ACC: 0.866<br>ACC: 0.865 |
| Lin et al. (2020) | 2020 | GCN | Propagation | Textual | Context-based | Supervised | rumor | Twitter15<br>Twitter16 | ACC: 0.856<br>ACC: 0.881 |
| Ren et al. (2020) | 2020 | GAT | Heterogeneous | Textual, Profile | Content-based | Semi- supervised | Fake news | PolitiFact<br>BuzzFeed | ACC: 0.615<br>ACC: 0.735 |
| Huang et al. (2020) | 2020 | GAT | Heterogeneous | Textual, Profile, Temporal | Content-based | Supervised | rumor | Twitter15<br>Twitter16 | ACC: 0.911<br>ACC: 0.924 |
| Ke et al. (2020) | 2020 | GCN | Propagation | Textual, Profile | Context-based | Supervised | rumor | Weibo | ACC: 0.969 |
| Nguyen et al. (2020) | 2020 | GraphSage | Heterogeneous | Textual, Profile | Hybrid | Supervised | Fake news | Twitter | AUC: 0.752 |

**Table 1** (continued)

| Ref. | Year | GNN Type | Graph | Features | Approach | Setting | Disinformation Type | Dataset | Performance |
|---|---|---|---|---|---|---|---|---|---|
| Lu and Li (2020) | 2020 | GCN | Similarity | Textual, Profile | Hybrid | Supervised | Fake news | Twitter15<br>Twitter16 | ACC: 0.877<br>ACC: 0.908 |
| Song et al. (2021) | 2021 | GAT<br>GCN | Propagation | Textual, Temporal | Context-based | Supervised | Fake news | Weibo<br>FakeNewsNe | ACC: 0.968<br>ACC: 0.935 |
| Yuan et al. (2021) | 2021 | GAT | Similarity | Textual, Visual | Content-based | Supervised | Fake news | Twitter<br>Weibo | Macro-F1: 0.883<br>Macro-F1: 0.952 |
| Wei et al. (2021) | 2021 | GCN | Propagation | Textual | Context-based | Supervised | Rumor | Twitter15<br>Twitter16<br>PHEME | ACC: 0.892<br>ACC: 0.915<br>ACC: 0.715 |
| Jiang et al. (2021) | 2021 | GAT | Propagation | Textual | Context-based | Supervised | Rumor | Twitter15<br>Twitter16 | ACC: 0.851<br>ACC: 0.883 |
| Choi et al. (2021) | 2021 | GCN | Propagation | Textual | Context-based | Supervised | rumor | Twitter15<br>Twitter16<br>Weibo | ACC: 0.827<br>ACC: 0.836<br>ACC: 0.936 |

**Table 2** GNN-based disinformation detection methods (Part 2 - continuation of Table 1). ACC stands for accuracy

| Ref. | Year | GNN Type | Graph | Features | Approach | Setting | Disinformation Type | Dataset | Performance |
|---|---|---|---|---|---|---|---|---|---|
| Ran et al. (2022) | 2022 | GAT | Heterogeneous | Textual, Profile, Temporal | Hybrid | Supervised | rumor | Twitter15<br>Twitter16 | ACC: 0.908<br>ACC: 0.916 |
| Song et al. (2022) | 2022 | GAT | Propagation | Textual | Context-based | Supervised | Fake news | Weibo<br>FakeNewsNet<br>Twitter | ACC: 0.957<br>ACC: 0.922<br>ACC: 0.899 |
| Liu et al. (2022) | 2022 | GAT | Heterogeneous | Textual, Profile, Comments | Hybrid | Supervised | rumor | Weibo<br>Twitter15<br>Twitter16 | ACC: 0.944<br>ACC: 0.905<br>ACC: 0.902 |
| Wei et al. (2022) | 2022 | GCN | Propagation | Textual | Context-based | Supervised | rumor | Twitter15<br>Twitter16<br>PHEME | ACC: 0.901<br>ACC: 0.908<br>ACC: 0.694 |
| Zheng et al. (2022) | 2022 | GAT | Heterogeneous | Textual, Visual, Comments | Hybrid | Supervised | rumor | PHEME<br>Weibo | ACC: 0.887<br>ACC: 0.889 |
| Inan (2022) | 2022 | GAT | Heterogeneous | Textual, Profile, Comments | Content-based | Supervised | Fake news | PolitiFact<br>GossipCop | ACC: 0.874<br>ACC: 0.802 |
| Weizhi et al. (2022) | 2022 | GCN | Similarity | Textual, Profile | Content-based | Supervised | Fake news | Snopes<br>PolitiFact | Macro-F1: 0.800<br>Macro-F1: 0.691 |
| Paraschiv et al. (2022) | 2022 | GCN<br>GAT<br>GraphSage | Propagation | Textual, Profile, Comments | Hybrid | Supervised | Fake news | US Elections dataset | ACC: 0.867 |
| Xu et al. (2022) | 2022 | GCN | Propagation | Textual | Hybrid | Supervised | rumor | Weibo<br>CED | ACC: 0.957<br>ACC: 0.882 |
| Yang et al. (2023) | 2023 | GAT | Propagation | Textual, Temporal | Context-based | Supervised | rumor | PHEME<br>Weibo | ACC: 0.882<br>ACC: 0.972 |
| Cui et al. (2023) | 2023 | GGNN | Similarity | Textual, Semantic | Content-based | Supervised | Fake news | LIAR<br>Constraint<br>Twitter15<br>Twitter16 | ACC: 0.868<br>ACC: 0.918<br>ACC: 0.946<br>ACC: 0.968 |
| Thota et al. (2023) | 2023 | GCN | Propagation | Textual, Profile | Context-based | Supervised | rumor | Twitter<br>Weibo | ACC: 0.800<br>ACC: 0.911 |

- Most methods are based on supervised learning, where only a few methods harness the potential of GNNs in semi-supervised learning.

## 6 Datasets

Online information can be collected from a wide range of sources, including news homepages, search engines, and social networking websites. However, manually determining the accuracy of this information poses a significant challenge and typically necessitates annotators with ample knowledge and expertise in the corresponding domain. Thus, the absence of a standardized dataset for disinformation poses a challenge (Shu et al. 2017).

In this section, we introduce ten known datasets that have been recently used for disinformation detection in GNN-based methods. Besides these datasets, there exist several other datasets in the literature. A larger list of available datasets for disinformation detection can be found in Phan et al. (2023). We may categorize these ten datasets based on domain, type of content, number of labels, size and platform. A concise comparison is presented in Table 3. The availability of features (textual, visual etc) is a crucial factor when selecting a dataset. Furthermore, since most of GNN-based methods rely on processing the propagation graph, the availability of the propagation graph within the dataset is an important factor in dataset selection. We discuss the availability of these information in the datasets, in the "Content" column. Some other details of the datasets are as follows:

- LIAR[2]: a decade-long collection of 12.8K manually labeled short statements, which is compiled from politifact.com and encompasses various contexts. It offers comprehensive analysis reports and provides source document links for each case.
- BuzzFeed News[3]: the dataset comprises 2,283 news articles related to politics, which is collected from Facebook.
- Yelp[4]: a subset of Yelp's businesses, reviews and user data, specifically focusing on hotel and restaurant reviews. It includes both filtered reviews labeled as "spam" and recommended reviews deemed "legitimate" by Yelp.
- Reddit[5]: this dataset is derived from a popular news aggregation site where users continuously share and comment on a vast amount of fresh Internet content.
- Fakeddit[6]: a multimodal fake news dataset, which includes images, comments and metadata news. It consists of over 1 million samples from multiple categories of fake news.
- PHEME[7]: a dataset which contains a collection of Twitter rumors and non-rumors posted during breaking news. The dataset includes rumors related to 9 events, with

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

[3] https://data.world/buzzfeednews.

[4] https://www.yelp.com/dataset.

[5] https://www.reddit.com/.

[6] https://paperswithcode.com/dataset/fakeddit.

[7] https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078.

each rumor annotated with its corresponding veracity value, categorized as either True, False, or Unverified.
- Twitter15 and Twitter16[8]: two datasets that contain the tweets and retweets from a thousand news articles published in 2015 and 2016.
- Sina Weibo[9]: a dataset collected based on the claims reported on www.weibo.com, where each claim contains text, image URL, response and so on.
- FakeNewsNet[10]: it contains two datasets collected using ground truths from Politifact and Gossipcop. The first collection was acquired from FakeNewsNet repository, which contains news of famous people, fact-checked by the GossipCop website.

We provide examples from two popular datasets, LIAR and Twitter15, in Table 4 to help readers comprehend the content and labeling of the datasets more clearly.

## 7 Open problems

The literature reviewed in this paper highlights notable advances in addressing the detection and mitigation of disinformation, as well as endeavors to enhance the reliability and trustworthiness of online content. However, there remain numerous challenges that have yet to be adequately addressed. In the following, we discuss some of these challenges, that can form interesting directions for future work.

*Multiclass classification*

Most of the current studies categorize information into two classes: fake or real. Several methods have been developed, boasting high accuracy rates of over 90%, for binary classification purposes. However, relying solely on binary classification as a criterion for determining the accuracy of information may not be adequate. This is because in many cases, the content cannot be strictly categorized as either entirely true or false (Moradi and Chehreghani 2023). Instead, certain portions of the information are true while others are false. Hence, considering the option of classifying information into multiple classes or utilizing a credit score system can be a more effective approach. However, only a limited number of researchers have dedicated their efforts to multiclass classification and as a result, in this setting existing algorithms suffer from relatively low accuracy rates, typically below 50%.

*Intent detection*

The majority of existing research primarily concentrates on detecting the authenticity of content while disregarding the underlying intent or purpose behind the content. In general, false information is often created with the specific intention of deceiving audiences. Occasionally, false information may take the form of satire and be published without any malicious intent. Therefore, a more precise classification of information which is done based on intent, can be particularly effective in identifying misinformation from disinformation.

[8] https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0.

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

[10] https://github.com/KaiDMML/FakeNewsNet.

**Table 3** Disinformatiom datasets

| Dataset | Domain | Content | Labels | Size | Platform |
|---|---|---|---|---|---|
| LIAR(Wang 2017) | politics | text | 6:<br>{true, mostly-true,<br>half-true, barely-true,<br>false, pants-fire} | 12,836 | politifact.com |
| BuzzFeed News(Shu et al. 2017) | politics | text | 4:<br>{mostly true,<br>mixture of true and false,<br>mostly false,<br>no factual content} | 2282 | facebook |
| Yelp(Barbado et al. 2019) | technology | text | 2:<br>{fake,<br>real} | 45,954 | yelp.com |
| Reddit(Shang et al. 2020) | society | text | 2:<br>{fake,<br>real} | 2780 | reddit.com |
| Fakeddit(Nakamura et al. 2019) | politics,<br>society | text,<br>image,<br>videos | 6:<br>{True,<br>Satire/Parody,<br>Misleading Content,<br>Imposter Content,<br>False Connection,<br>Manipulated Content} | 795,108 | reddit.com |
| PHEME(Zubiaga et al. 2016) | politics,<br>society | text,<br>propagation graph | 4:<br>{True,<br>False,<br>Unverified,<br>non-rumor} | 6425 | twitter |
| Twitter15(Ma et al. 2017) | society | text,<br>propagation graph | 4:<br>{True,<br>False,<br>Unverified,<br>non-rumor} | 1490 | twitter |

**Table 3** (continued)

| Dataset | Domain | Content | Labels | Size | Platform |
|---|---|---|---|---|---|
| Twitter 16(Ma et al. 2017) | society | text, propagation graph | 4: {True, False, Unverified, non-rumor} | 818 | twitter |
| Sina Weibo(Ma et al. 2016) | society | text, images, propagation graph | 2: {fake, real} | 4664 | weibo |
| FakeNewsNet(Shu et al. 2018) | politics, celebrities | text, images, propagation graph | 2: {fake, real} | 23,901 | politifact.com gossipcop.com |

**Table 4** Examples of records in the LIAR and Twitter15 datasets

| Dataset | Sample | Label |
|---|---|---|
| LIAR | Walking Dead: In the case of a catastrophic event, the Atlanta-area offices of the Centers for Disease Control and Prevention will self-destruct. | pants-fire |
| LIAR | Donald Trump: "NATO is opening up a major terror division. ...Im sure Im not going to get credit for it, but that was largely because of what I was saying and my criticism of NATO." | false |
| LIAR | Robin Vos: "The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades." | true |
| LIAR | Republican Party Texas: "Jim Dunnam has not lived in the district he represents for years now." | barely-true |
| LIAR | Scott Surovell: "When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration." | half-true |
| LIAR | Barack Obama: "Hillary Clinton agrees with John McCain "by voting to give George Bush the benefit of the doubt on Iran." | mostly-true |
| Twitter15 | rip elly may clampett: so sad to learn #beverlyhillbillies star donna douglas has passed away | true |
| Twitter15 | seriously? racist mcdonald's sign is obviously a hoax. | false |
| Twitter15 | an open letter to trump voters from his top strategist-turned-defector URL via @xojanedotcom | unverified |
| Twitter15 | brandon marshall visits and offers advice, support to brother of fallen hero zaevion dobson | non-rumor |

*Gold standard datasets*

The establishment of comprehensive and reliable gold standard datasets is highly essential in this field. Currently, a significant portion of research is conducted using customized datasets, which may limit the generalizability and comparability of the results. The scarcity of publicly available large-scale datasets poses a challenge in conducting solid benchmark comparisons between different algorithms. Furthermore, it is worth noting that the datasets currently available in this field are predominantly developed in the English language. This linguistic bias restricts the applicability and generalizability of algorithms and models to other languages, cultures, and contexts. Indeed, since the issue of fake news is a global problem, it is crucial to allocate more resources towards developing datasets and research materials in various languages. By expanding the coverage of different languages, we can effectively address the global nature of misinformation and disinformation, providing more comprehensive solutions and insights across different regions and cultures. A multilingual approach would enhance our ability to combat fake news on a global scale.

*Explainability*

Explainability has emerged as a prominent research area in artificial intelligence and machine learning. Unfortunately, most of false information detection models are not

explainable. Nonetheless, incorporating explainable and transparent outputs can prove highly effective in bolstering the reliability of these models. To achieve explainability in disinformation detection models, various approaches can be adopted. These include mining social feedback, such as detecting users' stances and emotions, utilizing more explainable neural models such as graph neural networks (Chehreghani 2022), and leveraging interdisciplinary research such as incorporating psychological theories.

*Unsupervised and semi-supervised learning*

The majority of existing approaches in this field are supervised, relying on labeled datasets for training machine learning models. Indeed, acquiring a reliable dataset of labeled disinformation is a time-consuming and costly process. It involves meticulous analysis of claims, evidence, contextual information and cross-referencing with credible sources. Therefore, developing unsupervised or semi-supervised models that can learn from unlabeled data may mitigate the difficulty of obtaining and labeling data.

*Cross-platform analysis*

Typically, individuals have user accounts on multiple social networks through which they share content. In such cases, identifying the original source of the content can pose a certain level of difficulty. This issue has rendered cross-platform detection a crucial challenge when it comes to tracking and combating disinformation across social networks.

*Early detection*

As fake news continues to spread, there is an increased likelihood of people trusting and believing it. Hence, early detection plays a pivotal role in combating disinformation as it can help prevent its widespread dissemination. However, in order to effectively detect disinformation, most of existing approaches require a certain time lapse after the spread of each content.

*Multilingual analysis*

Currently, the majority of existing approaches are primarily monolingual and predominantly focused on the English language. Regrettably, many other popular and regional languages have not yet been adequately considered in these approaches. Multilingual analysis of disinformation offers a more comprehensive understanding of the phenomenon, enabling the identification of non-varying features that transcend language boundaries. Such an approach can be highly beneficial for early detection of disinformation, enabling proactive measures to counter its spread across different languages.

*Cross-domain analysis*

Most of existing approaches tend to concentrate on a single specific method to detect deception, focusing on aspects such as content, propagation patterns, or stylistic characteristics. Cross-domain analysis, encompassing various aspects such as topic, website, language, images and URLs in conjunction, proves valuable in identifying invariant features and facilitating accurate detection of disinformation. By incorporating multiple dimensions and considering the interplay of these factors, more robust and comprehensive disinformation detection systems can be developed.

*Dynamic networks*

In many existing methods, particularly in propagation and network-based approaches, an assumption is that the network structure remains static and unchanged. However, in reality, the structure of social networks is dynamic and undergoes changes over time (Chehreghani 2021). Hence, it is crucial to consider this dynamism when detecting disinformation, as the evolving nature of social network structures can significantly impact the spread and identification of deceptive information.

*Visual features*

In the literature, content-based methods predominantly rely on textual content analysis, with limited attention given to the utilization of visual features. However, it is important to note that social networks consistently feature visual content, such as images or videos, alongside textual content. Furthermore, in recent years, there has been a rise in the development of sophisticated tools capable of manipulating and generating fake images and videos. Hence, extracting visual features from images or videos can serve as crucial indicators for detecting false information and therefore, they should be considered alongside textual content in disinformation detection efforts.

*Hybrid models*

A conclusion of our study is that most effective approaches are hybrid methods that use both content-based and context-based features, during the classification. Hybrid approaches that can simultaneously model different aspects of disinformation, such as text, visual features, propagation patterns and stances, despite having issues such as generating more complex models, data availability and selecting proper features, might perform better for detecting disinformation.

## 8 Conclusions

In this paper, we reviewed various aspects of disinformation detection, with a focus on methods that are based on graph neural networks (GNNs). First, we defined disinformation and listed several forms of it, including fake news, rumors and hoaxes. Then, given the primary focus of this paper on exploring the efficiency of GNNs in identifying disinformation, we provided a concise overview of these models and scrutinized their distinctive attributes. Subsequently, we focused on various features of disinformation that can be utilized for detection, encompassing content-based and context-based features. Next, we thoroughly examined the methodologies employed for detecting disinformation, considering two distinct perspectives: the algorithmic approach and the features utilized in the process. Finally, we discussed challenges and open problems in the literature, which can serve as insights and direct future research endeavors.

**Author contributions** Batool Lakzaei: conceived ideas, analyzed methods and models, wrote paper. Mostafa Haghir Chehreghani: conceptualized work, wrote paper, supervised work. Alireza Bagheri: supervised work.

**Funding** Not applicable.

**Data availability** Not applicable.

**Code availability** Not applicable.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Ethical approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References

Afroz S, Brennan M, Greenstadt R (2012) Detecting hoaxes, frauds, and deception in writing style online. In: 2012 IEEE symposium on security and privacy, pp 461–475. IEEE

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–36

Alzanin SM, Azmi AM (2018) Detecting rumors in social media: a survey. Proc Comput Sci 142:294–300

Asghari S, Chehreghani MH, Chehreghani MH (2022) On using node indices and their correlations for fake account detection. In: IEEE international conference on big data, big data 2022, Osaka, Japan, December 17–20, 2022, pp 5656–5661

Autef A, Matton A, Romain M (2020) Fake news detection using machine learning on graphs. Technical Report

Azarijoo B, Salehi M, Najari S (2023) A meta path-based approach for rumor detection on social media. arXiv preprint arXiv:2301.04341

Bai N, Meng F, Rui X, Wang Z (2021) Rumour detection based on graph convolutional neural net. IEEE Access 9:21686–21693

Baly R, Karadzhov G, Alexandrov D, Glass J, Nakov P (2018) Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:1810.01765

Barbado R, Araque O, Iglesias CA (2019) A framework for fake review detection in online consumer electronics retailers. Inf Process Manag 56(4):1234–1244

Benamira A, Devillers B, Lesot E, Ray AK, Saadi M, Malliaros FD (2019) Semi-supervised learning and graph neural networks for fake news detection. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 568–569. IEEE

Bian T, Xiao X, Xu T, Zhao P, Huang W, Rong Y, Huang J (2020) Rumor detection on social media with bi-directional graph convolutional networks. In: The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAAI 2020, the tenth aaai symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020, pp 549–556

Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci 497:38–55

Briscoe EJ, Appling DS, Hayes H (2014) Cues to deception in social media communications. In: 2014 47th Hawaii international conference on system sciences, pp 1435–1443. IEEE

Brody S, Alon U, Yahav E (2021) How attentive are graph attention networks? arXiv preprint arXiv:2105.14491

Carvalho C, Klagge N, Emanuel Moench (2011) The persistent effects of a false news shock. J Empir Finance 18(4):597–615

Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the 20th international conference on World wide web, pp 675–684

Chehreghani MH (2021) Dynamical algorithms for data mining and machine learning over dynamic graphs. WIREs Data Mining Knowl Discov 11(2):1393

Chehreghani MH (2022) Half a decade of graph convolutional networks. Nat Mach Intell 4(3):192–193

Chen Y, Conroy NJ, Rubin VL (2015) Misleading online content: recognizing clickbait as" false news". In: Proceedings of the 2015 ACM on workshop on multimodal deception detection, pp 15–19

Chen Y-C, Liu Z-Y, Kao H-Y (2017) Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp 465–469

Chen Y-C, Wu SF (2018) Fakebuster: a robust fake account detection by activity analysis. In: 2018 9th international symposium on parallel architectures, algorithms and programming (PAAP), pp 108–110. IEEE

Chen GY, Li ZN, Yinfeng L, Yingrong Q, Jinghua P, Yuhan Q, Jianxin C, Depeng J, Xiangnan H et al (2023) A survey of graph neural networks for recommender systems: challenges, methods, and directions. ACM Trans Recomm Syst 1(1):1–51

Choi J, Ko T, Choi Y, Byun H, Kim C (2021) Dynamic graph convolutional networks with attention mechanism for rumor detection on social media. PLoS ONE 16(8):e0256039

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2014) A fake follower story: improving fake accounts detection on twitter. IIT-CNR, technical report TR-03

Cui B, Ma K, Li L, Zhang W, Ji K, Chen Z, Abraham A (2023) Intra-graph and inter-graph joint information propagation network with third-order text graph tensor for fake news detection. Appl Intell 1:1–18

Dai E, Zhao Ti, Zhu H, Xu J, Guo Z, Liu H, Tang J, Wang S (2022) A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. arXiv preprint arXiv:2204.08570

Damian F (2022) Rpasdt-rumour rpasdt-rumor propagation and source detection toolkit. SoftwareX 17:100988

Di Pietro R, Cresci S, Petrocchi M, Spognardi A, Tesconi M (2013) Fake accounts detection on twitter. Consiglio Nazionale delle Ricerche, pp 1–13

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, Xiang Zhang (2020) Parameterized explainer for graph neural network. Adv Neural Inf Process Syst 33:19620–19631

Elyassami S, Alseiari S et al (2022) Fake news detection using ensemble learning and machine learning algorithms. Combating fake news with computational intelligence techniques. Springer, Berlin, pp 149–162

Feng Q, Chengyue G, Karishma S, Yan L (2018) Neural user response generator: fake news detection with collective user intelligence. In IJCAI 18:3834–3840

Giasemidis G, Singleton C, Agrafiotis I, Nurse JRC, Pilgrim A, Willis C, Greetham DV (2016) Determining the veracity of rumours on twitter. International conference on social informatics. Springer, Berlin, pp 185–205

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: International conference on machine learning, pp 1263–1272. PMLR

Groza A (2023) Towards detecting fake news using natural language understanding and reasoning in description logics. In: Measuring ontologies for value enhancement: aligning computing productivity with human creativity for societal adaptation: first international workshop, MOVE 2020, Virtual Event, October 17–18, 2020, revised selected papers, pp 57–72. Springer

Guo Q, Qiu X, Xue X, Zhang Z (2021) Syntax-guided text generation via graph neural network. Sci China Inf Sci 64:1–10

Gupta M, Zhao P, Han J (2012) Evaluating event credibility on twitter. In: Proceedings of the 2012 SIAM international conference on data mining, pp 153–164. SIAM

Gupta A, Lamba H, Kumaraguru P, Joshi A (2013) Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web, pp 729–736

Hamed SK, Ab Aziz MJ, Ridzwan YM (2023) Fake news detection model on social media by leveraging sentiment analysis of news content and emotion analysis of users comments. Sensors 23(4):1748

Hamilton W, Ying Z, Leskovec J (2021) Inductive representation learning on large graphs. Adv Neural Inf Process Syst 30:1

Hande A, Puranik K, Priyadharshini R, Thavareesan S, Chakravarthi BR (2021) Evaluating pretrained transformer-based models for covid-19 fake news detection. In: 2021 5th international conference on computing methodologies and communication (ICCMC), pp 766–772. IEEE

Han Y, Karunasekera S, Leckie C (2020) Graph neural networks with continual learning for fake news detection from social media. arXiv preprint arXiv:2007.03316

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp 770–778. IEEE Computer Society

Himdi H, Weir G, Assiri F, Al-Barhamtoshy H (2022) Arabic fake news detection based on textual analysis. Arab J Sci Eng 1:1–17

Hongyan R, Caiyan J, Pengfei Z, Xuanya L (2022) MGAT-ESM: multi-channel graph attention neural network with event-sharing module for rumor detection. Inf Sci 592:402–416

Hu G, Ding Y, Qi S, Wang X, Liao Q (2019) Multi-depth graph convolutional networks for fake news detection. CCF International conference on natural language processing and Chinese computing. Springer, Berlin, pp 698–710
Huang Q, Zhou C, Wu J, Wang M, Wang B (2019) Deep structure learning for rumor detection on twitter. In: 2019 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
Huang Q, Yu J, Wu J, Wang B (2020) Heterogeneous graph attention networks for early detection of rumors on twitter. In: 2020 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
Huang W, Wang Y, Yang J, Xu Y (2022) Stance detection based on user feature fusion. Comput Intell Neurosci. https://doi.org/10.1155/2022/5738404
Imaduwage S, Kumara PPNV, Samaraweera WJ (2022) Capturing credibility of users for an efficient propagation network based fake news detection. In: 2022 2nd international conference on computer, control and robotics (ICCCR), pages 212–217. IEEE
Inan E (2022) Zoka: a fake news detection method using edge-weighted graph attention network with transfer models. Neural Comput Appl 34(14):11669–11677
Jabardi M, Hadi AS (2020) Twitter fake account detection and classification using ontological engineering and semantic web rule language. Int J Mod Sci 6(4):404–413
Jiang J, Liu Q, Yu Q, Li G, Liu M, Liu C, Huang W (2021) Landscape-enhanced graph attention network for rumor detection. International conference on knowledge science, engineering and management. Springer, Berlin, pp 188–199
Jin X, Cao J, Jiang Y-G, Zhang Y (2014) News credibility evaluation on microblog with a hierarchical propagation model. In: 2014 IEEE international conference on data mining, pp 230–239. IEEE
Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, Feb 12–17, 2016, Phoenix, Arizona, USA, pp 2972–2978
Jingui H, Wenya W, Jingyi L, Shengchun W (2023) Text summarization method based on gated attention graph neural network. Sensors 23(3):1654
Kadek SI, Bayupati IPA, Made SAD (2022) Detection of fake news using deep learning CNN-RNN based methods. ICT Express 8(3):396–408
Karamchandani N, Franceschetti M (2013) Rumor source detection under probabilistic sampling. In: 2013 IEEE international symposium on information theory, pp 2184–2188. IEEE
Ke Y, Jiang H, Li T, Han S, Xiaofei W (2020) Data fusion oriented graph convolution network model for rumor detection. IEEE Trans Netw Serv Manag 17(4):2171–2181
Kini MGR (2022) Term frequency tokenization for fake news detection. Intell Cyber Phys Syst Internet Things 3(1):2023
Kishore V, Kumar M (2023) Enhanced multimodal fake news detection with optimal feature fusion and modified bi-lstm architecture. Cybern Syst 1:1–31
Kishwar A, Zafar A (2023) Fake news detection on Pakistani news using machine learning and deep learning. Expert Syst Appl 211:118558
Kumar KP, Geethakumari G (2014) Detecting misinformation in online social networks using cognitive psychology. HCIS 4(1):1–22
Lesce T (1990) Scan: deception detection by scientific content analysis. Law Order 38(8):3–6
Li S, Yang J, Liang G, Li T, Zhao K (2022) Sybilflyover: heterogeneous graph-based fake account detection model on social networks. Knowl-Based Syst 1:110038
Lin H, Zhang X, Fu X (2020) A graph convolutional encoder and decoder model for rumor detection. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), pp 300–306. IEEE
Liu Y, Wu YB (2018) Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp 354–361
Liu Y, Myle O, Naman G, Jingfei D, Mandar J, Danqi C, Omer L, Mike L, Luke Z, Veselin S (2019) Roberta: a robustly optimized Bert pretraining approach, 07
Liu X, Zhao Z, Zhang Y, Liu C, Yang F (2022) Social network rumor detection method combining dual-attention mechanism with graph convolutional network. IEEE Trans Comput Soc Syst 10:2350
Lu Y-J, Li C-T (2020) Gcan: graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648
Luca CG, Prashant S, Rocha LM, Johan B, Filippo M, Alessandro F (2015) Computational fact checking from knowledge networks. PLoS ONE 10(6):e0128193

Luo R, Zhao S, Cai Z (2021) Application of graph neural network in automatic text summarization. In: Theoretical computer science: 38th national conference, NCTCS 2020, Nanning, China, November 13–15, 2020, revised selected papers, pp 123–138. Springer

Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong K-F, Cha M (2016) Detecting rumors from microblogs with recurrent neural networks. In: Subbarao K (ed) Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, pp 3818–3824. IJCAI/AAAI Press

Ma J, Gao W, Wong K-F (2017) Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th annual meeting of the association for computational linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, Volume 1: Long Papers, pp 708–717

Mahmud FB, Rayhan MMS, Shuvo MH, Sadia I, Morol MK (2022) A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection. In: 2022 7th international conference on data science and machine learning applications (CDMA), pp 97–102. IEEE

Malhotra B, Vishwakarma DK (2020) Classification of propagation path and tweets for rumor detection using graphical convolutional networks and transformer based encodings. In: 2020 IEEE sixth international conference on multimedia big data (BigMM), pp 183–190. IEEE

Mansour D, Moosavi MR, Hadi SM (2022) DSS: a hybrid deep model for fake news detection using propagation tree and stance network. Expert Syst Appl 198:116635

Meel P, Vishwakarma DK (2020) Fake news, rumor, information pollution in social media and web: a contemporary survey of state-of-the-arts, challenges and opportunities. Expert Syst Appl 153:112986

Meike N, Jan T, Shreyasi P, Elisa N, Michelle P, Yasmin S, Jörg S, van Keulen M, Seifert C (2023) From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable ai. ACM Comput Surv 55(13s):1–42

Mihalcea R, Strapparava C (2009) The lie detector: explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 conference short papers, pp 309–312

Ming J, Yifan Z, Jian X, Min Z (2022) Gatsum: graph-based topic-aware abstract text summarization. Inf Technol Control 51(2):345–355

Mohammad SM, Parinaz S, Svetlana K (2017) Stance and sentiment in tweets. ACM Trans Internet Technol 17(3):1–23

Moradi M, Chehreghani MH (2023) Multilevel user credibility assessment in social networks. CoRR arxiv:abs/2309.13305

Mughaid A, Obeidat I, AlZu'bi S, Elsoud EA, Alnajjar A, Alsoud AR, Abualigah L (2023) A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. Multimed Tools Appl 1:1–26

Nakamura K, Levy S, Wang WY (2019) r/fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854

Nguyen V-H, Sugiyama K, Nakov P, Kan M-Y (2020) Fang: leveraging social context for fake news detection using graph representation. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 1165–1174

Oshikawa R, Qian J, Wang WY (2018) A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770

Paraschiv M, Salamanos N, Iordanou C, Laoutaris N, Sirivianos M (2022) A unified graph-based approach to disinformation detection using contextual and semantic relations. Proc Int AAAI Conf Web Soc Med 16:747–758

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Alessandro M, Bo P, Walter D (ed) Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp 1532–1543. ACL

Phan HT, Nguyen NT, Hwang D (2023) Fake news detection: a survey of graph neural network methods. Appl Soft Comput 1:110235

Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2017) A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638

Praseed A, Rodrigues J, Santhi TP (2023) Hindi fake news detection using transformer ensembles. Eng Appl Artif Intell 119:105731

Qazi M, Khan MUS, Ali M (2020) Detection of fake news using transformer model. In: 2020 3rd international conference on computing, mathematics and engineering technologies (iCoMET), pp 1–6. IEEE

Qazvinian V, Rosengren E, Radev D, Mei Q (2011) Rumor has it: identifying misinformation in microblogs. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp 1589–1599
Rai N, Kumar D, Kaushik N, Raj C, Ali A (2022) Fake news classification using transformer based enhanced lstm and bert. Int J Cogn Comput Eng 3:98–105
Ren Y, Wang B, Zhang J, Chang Y (2020) Adversarial active learning based heterogeneous graph neural network for fake news detection. In: 2020 IEEE international conference on data mining (ICDM), pp 452–461. IEEE
Rubin VL, Conroy NJ, Chen Y (2015) Towards news verification: Deception detection methods for news discourse. In: Hawaii international conference on system sciences, pp 5–8
Saquete E, Tomás D, Moreda P, Martínez-Barco P, Palomar M (2020) Fighting post-truth using natural language processing: a review and open challenges. Expert Syst Appl 141:112943
Schütz M, Schindler A, Siegel M, Nazemi K (2021) Automatic fake news detection with pre-trained transformer models. In: Pattern Recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, proceedings, Part VII, pp 627–641. Springer
Shah D, Zaman T (2012) Rumor centrality: a universal source detector. In: Proceedings of the 12th ACM sigmetrics/performance joint international conference on measurement and modeling of computer systems, pp 199–210
Shang L, Zhang Y, Zhang D, Wang D (2020) Fauxward: a graph neural network approach to fauxtography detection using social media comments. Soc Netw Anal Mining 10(1):1–16
Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol 10(3):1–42
Sharma A, Sharma MK, Dwivedi RK (2023) Exploratory data analysis and deception detection in news articles on social media using machine learning classifiers. Ain Shams Eng J 1:102166
Shi B, Weninger T (2016) Fact checking in heterogeneous information networks. In: Proceedings of the 25th international conference companion on World Wide Web, pp 101–102
Shiwen W, Fei S, Wentao Z, Xie X, Bin C (2022) Graph neural networks in recommender systems: a survey. ACM Comput Surv 55(5):1–37
Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newslett 19(1):22–36
Shu K, Wang S, Liu H (2017) Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:1712.07709, 8
Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2018) Fakenewsnet: a data repository with news content, social context and spatialtemporal information for studying fake news on social media. arXiv preprint arXiv:1809.01286
Shu K, Bhattacharjee A, Alatawi F, Nazer TH, Ding K, Karami M, Liu H (2020) Combating disinformation in a social media age. Wiley Interdiscipl Rev 10(6):e1385
Slovikovskaya V (2019) Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. arXiv preprint arXiv:1910.14353
Song C, Shu K, Bin W (2021) Temporally evolving graph neural network for fake news detection. Inf Process Manag 58(6):102712
Song C, Teng Y, Zhu Y, Wei S, Bin W (2022) Dynamic graph neural network for fake news detection. Neurocomputing 505:362–374
Sook L (2020) Academic library guides for tackling fake news: a content analysis. J Acad Librariansh 46(5):102195
Takayasu M, Sato K, Sano Y, Yamada K, Miura W, Takayasu H (2015) Rumor diffusion and convergence during the 3.11 earthquake: a twitter case study. PLoS ONE 10(4):e0121443
Thota NR, Sun X, Dai J (2023) Early rumor detection in social media based on graph convolutional networks. In: 2023 international conference on computing, networking and communications (ICNC), pp 516–522. IEEE
Varlamis I, Michail D, Glykou F, Tsantilas P (2022) A survey on the use of graph convolutional networks for combating fake news. Future Internet 14(3):70
Vedova MLD, Tacchini E, Moret S, Ballarin G, DiPierro M, de Alfaro L (2018) Automatic online fake news detection combining content and social signals. In: 2018 22nd conference of open innovations association (FRUCT), pp 272–279. IEEE
Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. Stat 1050:20
Vishwakarma DK, Meel P, Yadav A, Singh K (2023) A framework of fake news detection on web platform using convnet. Soc Netw Anal Mining 13(1):24

Vrij A (2005) Criteria-based content analysis: a qualitative review of the first 37 studies. Psychol Pub Policy Law 11(1):3
Wang WY (2017) " liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648
Wang Y, Qian S, Hu Jun, Fang Q, Xu C (2020) Fake news detection via knowledge-driven multimodal graph convolutional networks. In: Proceedings of the 2020 international conference on multimedia retrieval, pp 540–547
Wei L, Hu D, Zhou W, Yue Z, Hu S (2021) Towards propagation uncertainty: edge-enhanced Bayesian graph convolutional networks for rumor detection. arXiv preprint arXiv:2107.11934
Wei L, Dou H, Zhou W, Wang X, Hu S (2022) A neuro-fuzzy approach. IEEE transactions on neural networks and learning systems, modeling the uncertainty of information propagation for rumor detection
Weizhi X, Junfei W, Liu Q, Shu W, Wang L (2022) Evidence-aware fake news detection with graph neural networks. Proc ACM Web Conf 2022:2501–2510
Welling M, Kipf TN (2016) Semi-supervised classification with graph convolutional networks. In: International conference on learning representations (ICLR 2017)
Xu W, Chen H (2015) Scalable rumor source detection under independent cascade model in online social networks. In: 2015 11th international conference on mobile ad-hoc and sensor networks (MSN), pp 236–242. IEEE
Xu S, Liu X, Ma K, Dong F, Riskhan B, Xiang S, Bing C (2022) Rumor detection on social media using hierarchically aggregated feature via graph neural networks. Appl Intell 1:1–14
Xu K, Hu W, Leskovec J, Jegelka S (2018) How powerful are graph neural networks? arXiv preprint arXiv:1810.00826
Xu D, Ruan C, Körpeoglu E, Kumar S, Achan K (2020) Inductive representation learning on temporal graphs. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net
Yajie G, Shujuan J, Ning C, Chiu DKW, Na S, Chunrong Z (2023) Mdg: fusion learning of the maximal diffusion, deep propagation and global structure features of fake news. Expert Syst Appl 213:119291
Yang X, Lyu Y, Tian T, Liu Y, Liu Y, Zhang X (2021) Rumor detection on social media with graph structured adversarial learning. In: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, pp 1417–1423
Yang X, Ma H, Wang M (2023) Research on rumor detection based on a graph attention network with temporal features. Int J Data Warehousing Mining 19(2):1–17
Yeqing Y, Yongjun W, Peng Z (2023) A graph-based pivotal semantic mining framework for rumor detection. Eng Appl Artif Intell 118:105613
Ying R, He R, Chen K, Eksombatchai P, Hamilton WL, Leskovec J (2018) Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 974–983
Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J (2019) Gnnexplainer: generating explanations for graph neural networks. Adv Neural Inf Process Syst 1:32
You W, Agarwal PK, Chengkai L, Jun Y, Cong Y (2014) Toward computational fact-checking. Proc VLDB Endow 7(7):589–600
Yu Y (2018) Review of the application of machine learning in rumor detection. In: Proceedings of the 5th international conference on control engineering and artificial intelligence, pp 46–52
Yuan H, Tang J, Hu X, Ji S (2020) Xgnn: Towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 430–438
Yuan H, Zheng J, Qiongwei Y, Qian Y, Yan Z (2021) Improving fake news detection with domain-adversarial and graph-attention neural network. Decis Support Syst 151:113633
Zhang H, Fan Z, Zheng J, Liu Q (2012) An improving deception detection method in computer-mediated communication. J Netw 7(11):1811
Zhang C, Guo Q, Fu L, Ding J, Cao X, Long F, Wang X, Zhou C (2023) Finding the source in networks: an approach based on structural entropy. ACM Trans Internet Technol 23:125
Zheng J, Zhang X, Guo S, Wang Q, Zang W, Zhang Y (2022) MFAN: multi-modal feature-enhanced attention networks for rumor detection. In: Luc De R (ed) Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022, pp 2413–2419. ijcai.org
Zhiyuan W, Pi D, Chen J, Xie M, Cao J (2020) Rumor detection based on propagation graph neural network with attention mechanism. Expert Syst Appl 158:113595
Zhou X, Zafarani R (2020) A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv 53(5):1–40

Zhou J, Cui G, Shengding H, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M (2020) Graph neural networks: a review of methods and applications. AI open 1:57–81
Zhou X, Li J, Li Q, Zafarani R (2023) Linguistic-style-aware neural networks for fake news detection. arXiv preprint arXiv:2301.02792
Zubiaga A, Liakata M, Procter R, Hoi G, Wong S, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11(3):e0150989
Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R (2018) Detection and resolution of rumours in social media: a survey. ACM Comput Surv 51(2):1–36