

A Survey on False Information Detection: From A Perspective of Propagation on Social Networks

Kun Xie

The Chinese University of Hong Kong

xiekun@se.cuhk.edu.hk

Sibo Wang

The Chinese University of Hong Kong

swang@se.cuhk.edu.hk

Abstract

The proliferation of false information in the digital age has become a pressing concern, necessitating the development of effective and robust detection methods. This paper offers a comprehensive review of existing false information detection techniques, approached from a novel perspective that emphasizes the propagation characteristics of misinformation. We introduce a new taxonomy that categorizes these methods into homogeneous and heterogeneous propagation-based approaches, providing a deeper understanding of the varying scopes and complexities involved in information dissemination. For each category, we present a formal problem formulation, review commonly used datasets, and summarize state-of-the-art methods. Additionally, we identify several promising directions for future research, including the creation of a unified benchmark suite, exploration of diverse information modalities, and development of innovative rumor debunking tasks. By systematically organizing the vast array of current techniques, this work offers a clear overview of the research landscape, aiding researchers and practitioners in navigating this complex field and inspiring further advancements.

1 Introduction

In the digital era, social media platforms facilitate both the rapid dissemination of information and the spread of false content, such as rumors and fake news. These can have severe consequences, including misleading the public [33], inciting panic [68], and influencing elections [1, 24]. Detecting and debunking false information early is crucial to mitigate its harmful impacts on society. The way information spreads through social networks can reveal its veracity [4, 45]. False information often exhibits distinct propagation patterns, such as rapid spread and specific dissemination strategies. By analyzing these propagation patterns and characteristics, researchers can identify false content more effectively and obtain deeper insights into how misinformation spreads.

This study seeks to offer a comprehensive understanding of propagation-based false information detection, addressing a gap left by existing surveys. Previous works [16, 21, 57, 63] primarily focus on text content-based rumor detection using general deep learning techniques. Other studies [35, 48] emphasize the fusion of multiple modalities, integrating text and images to enhance detection capabilities. Additionally, some research targets specific aspects of false information detection. For example, [31] concentrates on early detection, while [2] examines detection within particular populations. Despite these contributions, the propagation characteristics of false information on social networks remain underexplored. This perspective offers valuable insights that extend beyond traditional content analysis, highlighting the dynamic nature of misinformation spread and its implications for detection strategies. In recent years, several surveys [14, 32, 46, 56] have summarized methods for false information detection utilizing Graph Neural Networks (GNNs) [17, 29]. However, these surveys predominantly concentrate on GNN techniques, lacking a comprehensive review and analysis from the perspective of propagation

information. While propagation information can indeed be modeled using graph topology, it is important to note that GNNs are not the sole technology employed in this domain. Architectures such as Transformers [3, 11, 66] and RNNs [8, 45] have also been extensively studied and applied. Consequently, focusing solely on GNN-based methods does not provide a complete overview of propagation-based approaches. Our work seeks to address this gap by offering a more holistic understanding of false information detection through the lens of propagation dynamics.

Specifically, we begin by reviewing the definition of false information and introducing the concept of information propagation on social networks. We categorize propagation based on its homogeneous or heterogeneous nature. Homogeneous propagation involves the dissemination of a source post through user interactions such as comments and retweets, forming a single-type propagation graph where all nodes represent posts. This type of propagation provides additional perspectives and debates that aid in assessing the veracity of the original news. In contrast, heterogeneous propagation encompasses a broader social context, including user metadata, semantic information, and cross-platform discussions. This rich context forms a heterogeneous social context graph with multiple node types, offering a comprehensive view of news propagation and enhancing veracity understanding. These two categories reflect different scopes and complexities of information propagation, each presenting unique insights and challenges for rumor detection.

Based on this categorization, we have reviewed false information detection methods that utilize both homogeneous and heterogeneous propagation. For each category, we first provide a formal problem formulation to help readers better understand the task. We then review the commonly used datasets and summarize existing methods. For methods based on homogeneous propagation, we categorize them from the perspective of propagation into three subcategories: 1) methods that model the dynamics of propagation, 2) methods that enhance robustness in the face of complex and noisy real-world propagation, and 3) methods that utilize Large Language Models (LLMs) to aid in understanding propagation. For methods based on heterogeneous propagation, we categorize them based on the types of social context they incorporate into two subcategories: 1) user-related context and 2) other context. Finally, we propose several feasible future research directions, including the development of an off-the-shelf unified benchmark suite and exploring alternative approaches to mitigate the impact of false information, to promote the advancement of this field. Figure 1 illustrates the taxonomy proposed in this paper.

The contributions of this work can be summarized as follows:

- *New Perspective.* This study reviews and synthesizes existing false information detection methods from the perspective of propagation, categorizing them into homogeneous and heterogeneous types. This approach provides insights into the varying scopes and complexities involved in information dissemination.
- *New Taxonomy.* We propose a novel taxonomy for propagation-based false information detection methods, initially dividing them based on their use of homogeneous or heterogeneous propagation. This taxonomy aids in systematically organizing the extensive array of existing methods.
- *Comprehensive Review and Summary.* This work offers a thorough and detailed review of existing methods for false information detection. For each propagation category, we present a formal problem formulation, review commonly used datasets, and summarize state-of-the-art methods. This comprehensive overview provides readers with a holistic understanding of the current research landscape in this field.

This paper is organized as follows: Section 2 introduces the concept of false information and commonly used terms in the field. Section 3 explores the characteristics and distinctions between homogeneous and heterogeneous propagation. Sections 4 and 5 provide comprehensive summaries of existing false

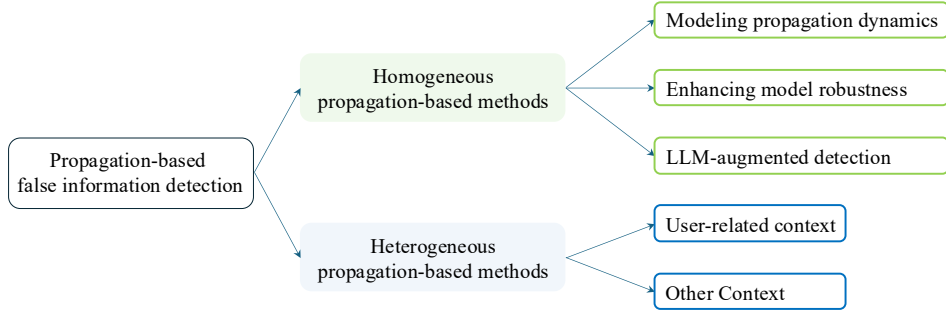


Figure 1: Taxonomy of propagation-based false information detection methods.

information detection methods based on homogeneous and heterogeneous propagation, respectively. Each section includes a formal problem formulation, a review of commonly used datasets, and an in-depth analysis of state-of-the-art methods. Finally, Section 6 discusses potential future research directions, highlighting opportunities for further advancement in this area.

2 False Information on Social Media

To comprehend false information on the web and social media, Kumar and Shah [30] systematically introduces this concept by categorizing it according to intent and knowledge content. Based on intent, false information can be classified as *misinformation*, which is created without the intention to mislead, and *disinformation*, which is deliberately produced to deceive and mislead readers [13]. From the perspective of knowledge content, false information can be divided into *opinion-based*, where a unique ground truth does not exist (as in the case of product reviews on e-commerce platforms), and *fact-based*, which involves falsehoods about entities that have a unique, verifiable ground truth [64].

In the field of false information detection, two terms are frequently encountered. The first term is *fake news* [57], which refers to deliberately false or verifiably inaccurate news content, falling within the category of disinformation under the aforementioned taxonomy of false information. The second term is *rumor*, defined as unverified and instrumentally relevant information statements in circulation [16, 43]. Unlike fake news, rumors may ultimately be proven true or false [93]. Despite their conceptual differences, these two terms are often used interchangeably in the context of social media false information detection, with their subtle distinctions being largely overlooked in previous studies [14]. Following this established practice in the literature, this survey reviews previous false information detection studies targeting both fake news and rumors.

False information can be disseminated through various types of media on social networks. In existing datasets, the common carriers include *micro-posts* on platforms such as Twitter¹ and sina Weibo², as well as published fake news *articles*. These carriers exhibit distinct characteristics: micro-posts are typically brief and colloquial, while news articles are longer and more structured, with separate components such as titles and content. While these structural differences are noteworthy, they are not the focus of our discussion. Instead, this paper emphasizes how false information propagates through social networks, regardless of its carrier type. In the next section, we present two primary approaches that existing works have employed to capture these propagation patterns.

¹<https://x.com/>

²<https://m.weibo.cn/>

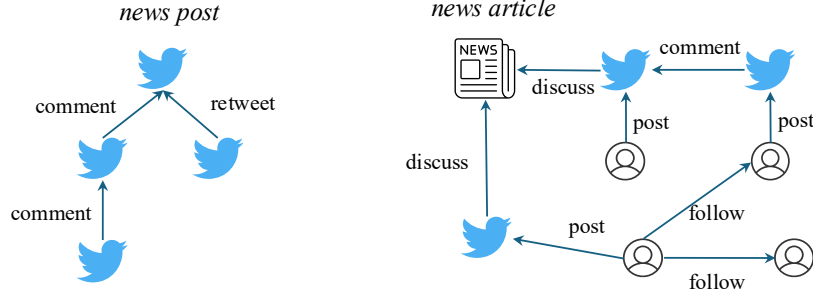


Figure 2: Illustration of homogeneous propagation (left) and heterogeneous propagation (right).

3 Information Propagation on Social Networks

The propagation of false information on social networks can be categorized into two distinct types based on the nature of propagation:

- **Homogeneous propagation** refers to the dissemination of a source post on social media, where other users interact with it through comments or retweets. These interactions, which may include expressing opinions on the source news or engaging in discussions with other comments, provide additional perspectives, supplementary information, and debates regarding the original news. The textual content of these retweets and comments offers further evidence for classifying the veracity of the original news. Moreover, the reply relationships between these discussions form a graph topology, linking the relevant texts together and further facilitating the prediction of veracity. We term this homogeneous propagation because it involves only a single type of propagation, forming a homogeneous propagation graph where all nodes represent posts on social media.
- **Heterogeneous propagation**, in contrast, encompasses the spread of information across broader social networks. Specifically, in addition to direct interactions with the original news, user metadata such as profiles and friendships, semantic information of the original news like keywords and domains, and related discussions of the news on other social platforms are all captured as social context. This rich social context provides a comprehensive view of the news propagation and discussion, enhancing the understanding of the news’s veracity. We refer to this as heterogeneous propagation because it involves multiple types of propagation forms. This heterogeneous propagation can be modeled with a heterogeneous social context graph, where nodes represent not just social media posts but also news articles, cross-platform discussions, and user accounts, while the edges connect different types of nodes, reflecting explicit and implicit information flow on social networks.

These two categories reflect the different scopes and complexities of information propagation on social networks, each offering unique insights and challenges for rumor detection. Figure 2 illustrates these two types of propagation. The left sub-figure displays the homogeneous propagation of a post on social media, with commenting and retweeting interaction posts forming the homogeneous propagation tree. Conversely, the right sub-figure demonstrates the propagation of a news article across complex social networks, where the news article is connected to related discussion posts on other social media platforms, and the user profiles and user friendship network are also collected, together forming a heterogeneous social context graph.

Based on these two distinct propagation mechanisms, we proceed to provide a comprehensive analysis of each approach. In Section 4, we present a formal problem formulation for homogeneous propagation-based rumor detection, followed by an overview of commonly used datasets and a systematic review of

existing methodologies. Similarly, Section 5 examines heterogeneous propagation-based false information detection, detailing its mathematical formulation, popular datasets, and a thorough survey of current research efforts.

4 Methods Based on Homogeneous Propagation

4.1 Problem Formulation

The task under this category is to classify a given post as either a rumor or a non-rumor, based on interactions occurring on the same social media platform, include commenting and retweeting the source post.

Formally, the propagation topology of the given source post can be modeled by a homogeneous graph, $G_p = (V_p, E_p, X_p)$, where the subscript p denotes the index of the source post. Specifically, the graph G_p is acyclic and tree-structured, as each comment or retweet is directed towards a single parent post. The node set V_p comprises all nodes in the graph G_p , with v_0 as the root representing the initial source post, and each v_i for $i \in \{1, \dots, |V_p| - 1\}$ representing a comment or retweet post. The edge set E_p includes all edges in the graph G_p , where each edge (v_i, v_j) signifies the comment or retweet relationship between two posts, i.e., post v_j comments on or retweets post v_i . The set X_p contains the text content of all posts, where i -th row of X_p denotes the text attributes of the corresponding node v_i . The source post, along with its direct propagation graph, is referred to as an event [4], and the dynamics of commenting and retweeting are sometimes called conversation threads [45].

The objective is to predict the veracity of the given post using both the text content and the direct propagation information G_p . The definition of rumor veracity varies across datasets. For instance, in the Twitter15 [44] and Twitter16 [44] datasets, posts are categorized into 1) non-rumors, 2) false rumors, 3) true rumors, and 4) unverified rumors. Conversely, in other datasets like PHEME5 [90, 92] and PHEME9 [91], posts are classified as 1) rumors and 2) non-rumors.

In a typical supervised learning framework, a function $f : \mathcal{G} \rightarrow \mathcal{Y}$ is learned to predict veracity labels for posts with propagation graphs, where \mathcal{G} represents the space of events, and \mathcal{Y} denotes the space of veracity labels. The function f is trained using a labeled training set \mathcal{G}_{train} , where the ground-truth veracity label y_p is available for each $G_p \in \mathcal{G}_{train}$. The learned function f is then applied to classify unlabeled events in the test set \mathcal{G}_{test} . Most existing propagation-based rumor detection studies follow this supervised learning approach, although some employ self-supervised learning or transfer learning paradigms, as will be discussed later.

Figure 3 illustrates an example event from the Weibo22 dataset [87], showcasing how false information about a flood spread on the Sina Weibo platform. The post claims that Yichang City in China experienced severe flooding, allegedly due to heavy rainfall and increased water discharge from the Three Gorges Dam. As this information circulated on the platform, users engaged through reposts and comments, reflecting the propagation threads and patterns of the source post. Notably, the comment posts contributed valuable, diverse perspectives and discussions, playing a crucial role in collectively verifying the authenticity of the source information. For instance, local users reported only minor waterlogging in subsequent comments and called for thorough fact-checking. By abstracting both the source post and its associated comments/reposts as nodes, the entire propagation process can be represented as a homogeneous graph, as depicted on the right side of the figure.

4.2 Datasets

We have compiled a collection of datasets commonly utilized in existing studies on homogeneous propagation-based detection of false information. Detailed descriptions of these datasets are provided

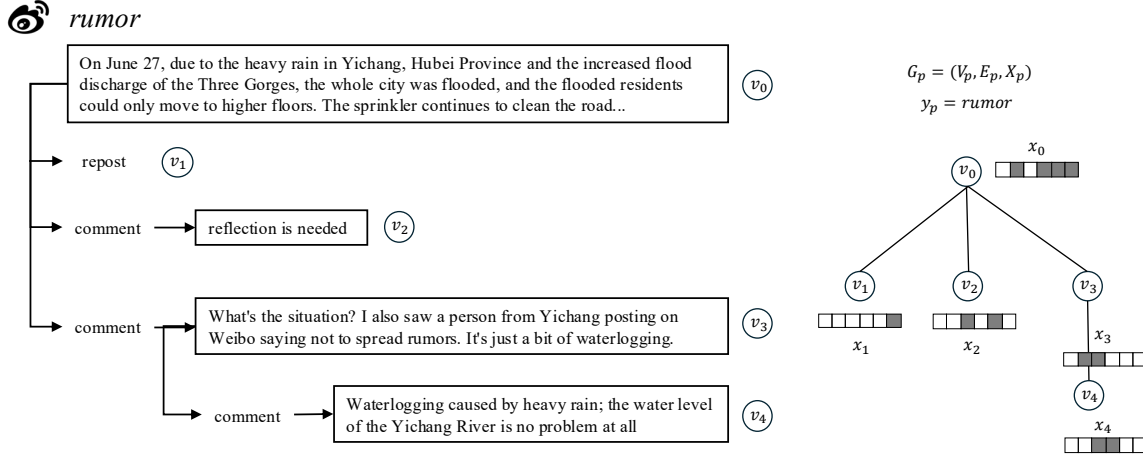


Figure 3: An example event from Weibo22 [87] containing false information related to a flood disaster, and the corresponding graph G_p modeling the propagation threads.

Table 1: Statistics of commonly used datasets, PHEME5 [90, 92], PHEME9 [91], Weibo-COVID19 [36], and Twitter-COVID19 [36], for direct propagation-based false information detection.

	PHEME5	PHEME9	Weibo-COVID19	Twitter-COVID19	Weibo	Weibo21	Weibo22
# Source posts	5,802	6,425	399	400	4,664	7,567	4,174
# Comment/retweet posts	103,212	105,354	26,687	406,185	3,805,656	184,775	961,962
Avg. # comment per event	18	16	67	1015	816	24	230
# Non-rumor	3,830	4,023	146	148	2,351	3,605	2,087
# Rumor	1,972	2,402	253	252	2,313	3,962	2,087

below:

- The **PHEME5** [90, 92] dataset was constructed by harvesting tweets from Twitter that pertain to newsworthy events with the potential to incite and propagate rumors. This dataset collected tweets based on five such events: *the Ferguson unrest*, *the Ottawa shooting*, *the Sydney crisis*, *the Charlie Hebdo shooting*, and *the Germanwings flight crash*. The collected tweets were then identified as rumors and non-rumors by journalists based on the source tweet contents and their associated comment threads.
- The **PHEME9** [91] dataset includes posts and propagation graphs derived from comments and reposts on Twitter. It encompasses tweets related to nine events, including breaking news likely to generate multiple rumors and known rumor events. Specifically, the dataset covers five sudden news events: *the Ferguson unrest*, *the Ottawa shooting*, *the Sydney crisis*, *the Charlie Hebdo shooting*, and *the Germanwings flight crash*. Additionally, it includes four specific rumors: *the Prince concert in Toronto*, *the Gurlitt collection*, *Putin's disappearance*, and *Michael Essien contracted Ebola*. Similar to PHEME5, journalists annotated the source posts for the collected tweets as either rumors or non-rumors, along with their propagation threads.

- The **Weibo** [43] dataset was constructed based on Sina Weibo, a large Chinese microblog platform. The dataset comprises a set of known rumors sourced from the Sina community management center³, documenting various instances of false information. An equivalent number of non-rumor events were gathered by randomly crawling posts not reported as rumors.
- The **Weibo21** [50] dataset is a multi-domain fake news dataset curated based on the Sina Weibo platform. It includes rumor events collected from reports by the Weibo Community Management Center, while non-rumor events were collected randomly and further verified by NewsVerify⁴. All news items were annotated across nine domains: *Science, Military, Education, Disasters, Politics, Health, Finance, Entertainment, and Society* using crowd-sourcing.
- The **Weibo22** [87] dataset was designed to address the issue of incomplete and untraceable posts in the previous Weibo and Weibo21 datasets. Specifically, Weibo22 provides original comment text and comprehensive author profile details. Events in the dataset are categorized as rumors or non-rumors based on information from the Weibo Community Management Center and the China Internet Joint Rumor Debunking Platform⁵. Over half of the posts in Weibo22 relate to the COVID-19 pandemic.
- The **Weibo-COVID19** and **Twitter-COVID19** [36] datasets were developed focusing on the COVID-19 pandemic. The Weibo-COVID19 dataset comprises rumor posts on Sina Weibo related to COVID-19, annotated by the Sina community management center, and non-rumor posts collected randomly. The Twitter-COVID19 dataset extends text-only claims from a COVID-19 rumor dataset [26] by incorporating propagation threads from Twitter. These datasets include micro posts in Chinese and English, respectively.
- The **Twitter15** and **Twitter16** [44] datasets extend previous text-only datasets by incorporating propagation trees. Twitter15 was constructed based on the framework proposed in [39], and Twitter16 was an extension of the dataset in [43]. Furthermore, the original binary labels (rumors and non-rumors) were refined to include four categories: non-rumors, false rumors, true rumors, and unverified rumors, in accordance with the veracity tags on articles from rumor debunking websites⁶.

Tables 1 and 2 present common statistics for nine popular datasets. Specifically, Table 1 provides statistics for seven datasets where news posts are classified into rumors and non-rumors, while Table 2 displays statistics for two datasets with news posts classified into four categories: non-rumors, false rumors, true rumors, and unverified rumors. It is important to note that the statistics of these datasets may vary over time due to the dynamic nature of social networks, where users frequently delete posts or deactivate their accounts. Consequently, this can lead to inconsistencies in dataset statistics across different studies.

4.3 Method Review

In this section, we review related works that detect false information on social media based on direct propagation threads. We categorize existing methods into three categories: 1) methods that focus on modeling the dynamics of propagation, 2) methods that focus on improving the robustness of propagation-based detection, and 3) methods that utilize large language models (LLMs) to augment propagation. We elaborate on the related methods from these aspects in the following sections.

³<https://service.account.weibo.com/>

⁴<https://www.newsverify.com/>

⁵<https://www.piyao.org.cn/>

⁶snopes.com and [Emergent.info](https://emergent.info)

Table 2: Statistics of commonly used datasets, Twitter15 [44] and Twitter16 [44], for direct propagation-based false information detection.

	Twitter15	Twitter16
# Source posts	1,490	818
# Comment/retweet posts	331,612	204,820
Avg. # comments per event	223	251
# Non-rumor	374	205
# False rumor	370	205
# True rumor	372	205
# Unverified rumor	374	203

4.3.1 Dynamics of Propagation

RvNN [45] and BiGCN [4] model the dynamics of propagation threads through two distinct directions: the bottom-up direction, which models rumor dispersion, and the top-down direction, which models rumor propagation. In the bottom-up tree, responsive nodes always point to the nodes they are responding to, resembling a citation network where a response acts like a reference. Conversely, the top-down structure aligns with the natural flow of information, where a link from node v_i to v_j indicates that v_j reads the information from v_i and responds to it. This setup simulates how information cascades from the source tweet to all its receivers. Based on these two directed tree structures, RvNN and BiGCN utilize recursive neural models [8] and graph neural networks [29] for rumor representation and classification, respectively. PLAN [27] and RDLT [83] further extend these directed relationships to pairwise interactions using learnable attention mechanisms. Specifically, PLAN employs the attention mechanism in a transformer network to capture long-range relationships among comments, while RDLT uses it to enhance focus on long-tail comments. DDGCN [60] models the propagation dynamics with a graph convolution network that captures structural information at different time stages, and then combines the representations from different stages with a temporal fusion unit. CT-RvNN [55] enhances the propagation model by incorporating comment time intervals as edge weights and constructs a coding tree [34] over the weighted propagation graph to extract essential structural information.

4.3.2 Model Robustness

In real-world scenarios, social media posts are often accompanied by unreliable responses, such as malicious comments designed to enhance the credibility of false information, and irrelevant noise comments unrelated to the source post. These factors significantly interfere with the accurate detection of rumors [42, 78, 91]. Therefore, enhancing the robustness of rumor detectors against noise, perturbations, and adversarial attacks, and developing models that can maintain high predictive performance in complex and noisy social media environments, is a crucial research direction. EBGCN [71] models comment uncertainty by adaptively adjusting edge weights in propagation graphs from a probabilistic perspective, which are then utilized in the aggregation of GCNs. FGCN [72] introduces a neuro-fuzzy method based on fuzzy theory [52] to adapt edge weights according to predefined membership. DCE-RD [86] designs a subgraph generation strategy to produce diverse counterfactual evidence, employing a Determinantal Point Process (DPP) based loss to enhance evidence diversity. The generated evidence is aggregated to provide robust veracity prediction. GARD [62] enhances detection robustness by capturing

semantic evolution information through self-supervised learning, conducting feature reconstruction from both local and global perspectives. EIN [22] addresses sensitivity to data quality by integrating epidemiological knowledge [6, 70], modeling propagation dynamics and user stance with susceptible states, and utilizing LLMs to annotate comment stances. Similarly, JS DRV [77] employs LLMs for post-level stance annotation and develops a reinforcement learning-based model for joint stance detection and rumor veracity prediction.

Contrastive learning, a self-supervised approach, ensures graphs maintain similar embeddings under different perturbations by emphasizing invariant features, thereby reducing sensitivity to noise and enhancing model robustness and generalization [7, 18]. This approach is also utilized to improve the robustness of rumor detectors. RDEA [19] pre-trains a GIN [75] encoder by maximizing mutual information between views under different perturbations, then fine-tunes the encoder for rumor veracity prediction. Node masking, edge dropping, and subgraph sampling operations are utilized for graph perturbation. RDCL [40] further proposes six perturbation methods under node-based and topology-based categories, maximizing consistency between two perturbed graphs of the same original graph, and minimizing the distance between perturbed and original graphs from the same class. A hard positive strategy also enhances the performance of the contrastive learning framework. GACL [61] utilizes contrastive learning to perceive differences between propagation graphs of the same and different classes, proposing an Adversarial Feature Transformation module to generate hard negative samples. CRFB [41] introduces a two-component (true-false) beta mixture model to distinguish true and false negative examples in contrastive learning, employing a CNN-based model to capture consistent and complementary information between two augmented propagation structures. RAGCL [9] observes the wide and shallow characteristics of propagation trees and proposes augmenting original graphs following three principles: exempting root nodes, retaining deep reply nodes, and preserving lower-level nodes in deep sections, based on which node centrality-based importance scores are incorporated to generate augmented views. FADE [85] improves representation quality with contrastive learning among augmented graphs under an adaptive strategy, mitigating event bias by subtracting event-only predictions.

Recent studies have demonstrated the effectiveness of adversarial training in improving robustness against noise and attacks [88]. AARD [59] considers both text content features and topology context when generating adversarial responses, utilizing the adversarial response generator to improve rumor detector robustness. GACL [61] also employs adversarial training to produce conflicting samples and hard negative samples, enhancing the effectiveness and robustness of the rumor detector.

Generation-based methods are also employed to enhance rumor detector robustness under certain conditions. GenFEND [51] focuses on alleviating exposure bias and enhancing comment diversity, adopting large language models (LLMs) as user simulators and comment generators. KPG [87] comprehensively considers both noisy propagation threads and spreading-limited rumors at early stages, proposing two interdependent modules to select key propagation graphs from generation-enhanced candidate graphs, utilizing a reinforcement learning framework to alternately update these modules. D² [74] aims to enhance early rumor detection by predicting possible diffusion paths based on limited propagation at early stages and users' social relationships.

Furthermore, some methods target improving transferability among rumors in different languages or topics. ACLR [36] overcomes domain and language restrictions via language alignment and a novel supervised contrastive training paradigm, developing an adversarial augmentation mechanism to enhance the robustness of low-resource rumor representation. To assess transferability under different test topics and languages, the Weibo-COVID19 dataset is utilized as the low-resource test set for evaluating proposed methods, while the Twitter15 and Twitter16 datasets serve as the well-resourced training set. Similarly, the Twitter-COVID19 dataset is used as the low-resource test set, and the Weibo dataset is employed as the well-resourced training set. T³RD [84] introduces test-time self-supervised learning to enhance rumor detection performance on low-resource datasets, performing graph- and node-level contrastive learning

as auxiliary tasks, and introducing feature alignment constraints to balance knowledge derived from the training set and test samples. FNDCD [15] employs a reweighting strategy based on classification confidence and propagation structure regularization to reduce domain-specific biases, enhancing the detection of unseen fake news with new topics and domains.

4.3.3 LLM-augmented Detection

In recent years, there has been a growing interest in leveraging Large Language Models (LLMs) for rumor detection. Despite their remarkable capabilities in text understanding and generation, LLMs face significant challenges when dealing with complex structured information, particularly in effectively processing intricate topologies. Specifically, LeRuD [38] designs specific prompts to guide the model’s focus on crucial clues such as writing styles, commonsense errors, and rebuttals or conflicts in comments. It also divides the entire propagation information into a Chain-of-Propagation to alleviate the cognitive burden on LLMs. ARG [20] prompts the LLM to generate multi-perspective rationales for rumors and utilizes smaller language models, such as BERT [11], to integrate and infer rumor veracity from these rationales. SePro [82] employs Graph Attention Networks [37] and community detection methods [53] to identify and extract key subgraphs from the complex information network, thereby simplifying the structure that the LLM needs to process.

5 Methods Based on Heterogeneous propagation

5.1 Problem Formulation

As introduced in Section 4.1, the task of rumor detection based on homogeneous propagation information involves predicting the veracity of the source post based on the commenting/retweeting interactions that form a propagation tree, making it a graph-level classification problem. In contrast, rumor detection with heterogeneous propagation incorporates a broader spectrum of information collected from social networks. This approach can be viewed as a node classification problem on heterogeneous graphs, where different types of nodes and edges represent various elements of the social context. Specifically, the social context typically includes: 1) users who published the source, comment, or retweet posts; 2) keywords extracted from the text content of the source news; and 3) semantically related posts discussing the same news event on other social platforms. To integrate all this information, a heterogeneous graph is constructed, with different types of nodes and edges representing the complex relationships within the social context.

Unlike rumor detection based on direct propagation information, which primarily focuses on the structure and dynamics of information spread, approaches with social context-based propagation introduce a higher level of complexity. Different methods in this category often adopt varied problem definitions, focusing on distinct aspects of the social context. To facilitate a comprehensive understanding, we abstract a foundational problem formulation that serves as a common ground for different variants of the problem.

Specifically, let S denote the set of news posts, T denote the set of comment/retweet posts interacting with the news posts in S , and U denote the set of users on the social network. The set \mathcal{U} includes users who directly interacted with the news posts (i.e., by publishing the news posts or replying to them with comments/retweets), as well as users indirectly associated with the news posts (i.e., having a friend relationship with direct users). These sets collectively form the different types of nodes in a heterogeneous graph, $V = S \cup T \cup U$. In this heterogeneous graph, the edges encompass various types of relationships, denoted by the edge set E . Specifically, edges between nodes in S and T indicate commenting or retweeting interactions. Edges between nodes in U and S represent users publishing news

posts $s_i \in S$, or commenting on/retweeting news posts $s_i \in S$, while edges between $u_i \in U$ and $t_i \in T$ represent u_i posting the comment/retweet post t_i . Additionally, edges between nodes in U capture social relationships, such as interactions or friendships. Thus, the heterogeneous graph can be represented as $G = (V, E)$. The goal of this task is to learn a prediction function $f : S \times \mathcal{G} \rightarrow \mathcal{Y}$, which classifies the news posts $s_i \in S$ into rumors (fake news) or non-rumors (real news) along with the heterogeneous social context propagation graph G , where \mathcal{Y} is the space of veracity labels. The learning process is based on a set of labeled news posts \mathcal{S}_{train} , and the learned function f is expected to accurately predict the veracity label of unlabeled news posts in \mathcal{S}_{test} : $\hat{y}_i = f(s_i, G)$. In some existing methods, only a subset of the node/edge types in the heterogeneous graph G are utilized to enhance the effectiveness of rumor detection, while other approaches employ additional information to refine the types and relationships within G .

5.2 Datasets

In this section, we introduce the common datasets used by false information detection methods that leverage heterogeneous propagation. In addition to the Twitter15, Twitter16, and Weibo datasets introduced in Section 4.2, the following datasets are also utilized:

- **Fakeddit** [49] is a large-scale benchmark dataset from Reddit⁷ that includes both text and image content. This dataset supports news classification at different granularities, including 2-class, 3-class, and 6-class classifications. Specifically, the 2-way classification determines whether a sample is fake or true. The 3-way classification distinguishes between completely true samples, direct quotes from propaganda posters, and fake samples. The 6-way classification includes the following labels: *True*, *Satire/Parody*, *Misleading Content*, *Imposter Content*, *False Connection*, and *Manipulated Content*.
- **GossipCop** [58] collects news from GossipCop⁸, a website that provides rating scores for entertainment stories on a scale from 0 to 10, reflecting the degree from fake to real. The dataset includes stories with scores under 5 as fake news, and collects related discussion posts on Twitter, along with their commenting/retweeting interactions, engaged user profiles, user friendship information, and spatial locations provided in user profiles.
- **PolitiFact** [58] gathers news content based on annotations from PolitiFact⁹, a fact-checking website where political news is evaluated as fake or real by journalists and domain experts. Similar to the GossipCop dataset, the PolitiFact dataset also includes semantically related tweet posts on Twitter, along with rich propagation information on social contexts, including both comment/retweet interactions and comprehensive user information associated with the posts.
- **FakeHealth** [10] is collected from the healthcare information review website, Health News Review¹⁰, and contains over 2000 news articles, 500k posts, and 27k user profiles, along with user networks.
- **Mc-Fake** [47] comprises labeled fake news from existing datasets and real news from reliable sources, covering five topics: *Politics*, *Entertainment*, *Health*, *Covid-19*, and *Syria War*. Related tweets, retweets, and replies, along with the corresponding users on the Twitter platform, are retrieved, forming a dataset with rich social contexts.

⁷<https://www.reddit.com/>

⁸<https://www.gossipcop.com/>

⁹<https://www.politifact.com/>

¹⁰<https://www.healthnewsreview.org/>

5.3 Method Review

In this section, we compile and review existing methods for false information detection that leverage heterogeneous propagation information on social networks. Unlike the methods discussed in Section 4, which focus on homogeneous propagation patterns, the methods reviewed here incorporate a broader range of contextual information available on social networks. Specifically, we classify these methods into two main categories based on the types of additional context they utilize: 1) user-related context and 2) other context.

5.3.1 User-Related Context

In this category, methods collect information about engaged users as social context, providing complementary evidence to assist in detecting false information. The user context not only provides connections across different news events, implying implicit information dissemination, but also reflects user preferences through user profiles and friendship relationships, offering useful clues for veracity classification. Specifically, GLAN [80] integrates user information to model local and global propagation on social networks using a heterogeneous graph, involving news posts, commenting posts, and the users publishing these posts. An attention mechanism is employed to enhance the representation of news posts by fusing the semantic information in the constructed graph. SMAN [81] further categorizes users into two types: 1) publishers, the authors of news posts, and 2) comment users who post comments on news posts. Additionally, SMAN annotates the credibility of a subset of users based on their historical publishing and reposting behaviors as supervised information, and jointly predicts rumor veracity and user credibility. FANG [54] follows this user type refinement and proposes an inductive graph learning framework [17] that produces representations effectively and efficiently. UPFD [12] is inspired by the confirmation bias theory, which suggests that the probability of users spreading fake news is related to their preferences, which can be extracted from users' historical posts. Specifically, UPFD models the target news and user historical post contents, extracts user-user interaction information, and fuses these endogenous preferences and exogenous context together. DUCK [65] extracts information from the user-user interaction network, the post-post commenting network, and conversation threads using Graph Attention Networks (GAT) [67], BERT [11] combined with GAT, and transformers [3]. PSIN [47] applies a divide-and-conquer strategy to model the heterogeneous relations and employs three variants of Graph Attention Networks to model post-post, user-user, and post-user subgraphs. An additional adversarial topic discriminator is adopted to learn topic-agnostic features for enhanced veracity prediction. DECOR [73] computes the number of common engaged users between news articles to form a news engagement graph. DECOR further reveals that the degree distributions of fake and real news in this engagement graph show a clear difference, and proposes a Degree-Corrected Stochastic Block model to enhance fake news detection. PSGT [89] focuses on user-user interaction relationships, proposing a noise-reduction self-attention mechanism based on the information bottleneck principle and a novel relational propagation graph as a position encoding for the graph Transformer.

Special learning settings, such as adversarial attack and transfer learning, are also explored. For instance, QSA-AC [69] considers attacking the user-news engagement network by injecting bots into the network to obfuscate GNN-based fake news detectors. QSA-AC follows a surrogate-based approach, leveraging the query process to enhance the effectiveness of the attack. Meanwhile, MMHT [79] utilizes user engagement and historical preference information for cross-domain fake news detection. Specifically, MMHT disentangles news content and user engagement from a macro perspective and disentangles veracity-relevant and veracity-irrelevant features from a micro perspective. These features are then utilized to facilitate more effective knowledge transfer.

5.3.2 Other Context

In addition to leveraging user information as social context, methods in this category broaden the scope of social context by incorporating various types of information. For instance, HDGCN [25] collects the temporal and domain information of news articles and proposes a neighbor sampling strategy along with a novel hierarchical attention mechanism to enhance news embedding and classification. DiHAN [5] constructs a heterogeneous graph that includes the core subjects of the news and their temporal information. It extracts meta-paths considering both node/edge types and the chronological interactions between news items, employing a hierarchical attention mechanism to integrate different types of meta-path-based temporal information. Temporal information is also utilized in DWAN [28] for train-test splitting, introducing a new approach where the model is trained on earlier news and tested on subsequent news. DWAN improves detection performance under this temporality-aware setting by using a Graph Structure Learning framework to reassign weights to existing edges based on a learned edge weight estimator. FinerFact [23] mines social context information through semantic relations to enable fine-grained reasoning. Specifically, FinerFact extracts keywords from news articles and posts on other social media platforms, using them as connectors between news, discussion posts, and the users who publish these posts. This forms a heterogeneous graph rich in social contextual information. The credibility of users is combined with the graph topology to reason about the veracity of rumors. Following this graph construction approach, SureFact [76] proposes a reinforced subgraph generation method and develops a Hierarchical Path-aware Kernel Graph Attention Network to reason about rumor veracity based on the generated subgraphs. This subgraph reasoning paradigm not only provides clear interpretability of the prediction results but also enhances generalization.

6 Future Work

6.1 Development of a Comprehensive Benchmark for False Information Detection

The field of false information detection currently lacks a practical, off-the-shelf benchmark that can standardize research and facilitate direct comparisons. This deficiency is particularly evident in three critical aspects: unified train-validation-test splitting, standardized evaluation process, and unified data platform.

Firstly, in existing studies, the train-test split is often conducted independently by different researchers, leading to non-comparable outcomes across studies and inconsistency with real-world scenarios. To address these challenges, we propose the development of a benchmark that includes commonly used datasets and provides a unified train-validation-test split. This split should be based on meaningful dimensions, such as topic or time, to better simulate real-world conditions and assess the robustness and generalizability of detection methods. For instance, splitting the dataset by topic or time stamp, with the training and test sets containing entirely different topics, can effectively evaluate a method’s ability to generalize across diverse subjects. This approach also allows researchers to analyze how generalization performance varies with different topics.

Secondly, another critical gap in current research is the lack of a unified evaluation process. Different methods often employ distinct evaluation codes and adopt various metrics, making it difficult to compare their performance directly. A standardized benchmark should include a consistent evaluation protocol, with clearly defined metrics and evaluation scripts. This will ensure that all methods are assessed under the same conditions, facilitating fair comparisons and promoting the development of more effective detection techniques.

Thirdly, the dispersion of datasets across different studies also poses challenges for researchers. A comprehensive, off-the-shelf benchmark platform that integrates multiple datasets would significantly

facilitate research in this area. Such a platform would provide a one-stop solution for researchers to access and work with standardized datasets, reducing the time and effort required to prepare data and allowing them to focus more on developing and testing new methods. This centralized approach would also promote collaboration and reproducibility, driving the field forward more efficiently.

6.2 Diversifying Information Carriers for False Information Detection

Current propagation-based false information detection primarily focuses on news or microblogs, where text is the dominant information carrier, occasionally supplemented by images. However, with the rapid growth of short video platforms, false information disseminated in video format has become a significant concern. The detection of false information in video content and the construction of relevant datasets should be prioritized in future research.

Short videos offer a unique set of challenges and opportunities for false information detection. Unlike text, which can be analyzed through natural language processing techniques, video content involves multiple modalities, including visual, auditory, and textual elements. Developing methods that can effectively integrate and analyze these multimodal features will be crucial for accurate detection. Additionally, creating datasets that capture the nuances of false information spread through videos will provide a valuable resource for researchers to develop and validate new detection algorithms. This diversification of information carriers will not only address the evolving landscape of information dissemination but also enhance the robustness and applicability of false information detection techniques.

6.3 Mitigating the Impact of False Information

Detecting and removing false information from social platforms (e.g., by deleting posts) is a straightforward approach to mitigate the impact of false information. However, this method has several drawbacks. First, it may mistakenly remove content that is not false information, thereby affecting user experience. Second, it does not address the root causes of why people believe and propagate false information. Finally, this method heavily relies on platform operations and may not comprehensively solve the problem of false information. Therefore, how to effectively reduce the impact of potentially false information on social networks is a critical issue that requires in-depth research.

One potential solution is to utilize bots on social platforms to disseminate corrective information. This approach can automatically push accurate information to users once false information is detected, thereby mitigating the spread of misinformation. However, several key aspects need to be considered for this method to be effective. The content of corrective information needs to be based on both the original information and the propagated information. This is because the propagated information reflects the community’s understanding of the original message, and generating corrective posts from this understanding can increase the likelihood of acceptance. Furthermore, the dissemination strategy for corrective information should be precisely tailored by integrating the propagation information of the original post and relevant social context information.

Despite its potential, this approach currently faces many challenges. The problem definition and evaluation of such a correction mechanism are still unclear. For example, how to measure the effectiveness of corrective information and determine its role in reducing the spread of false information remains an open question. Moreover, there is a lack of high-quality datasets for researching this issue. Existing datasets mostly focus on the detection of false information, while data related to the generation, dissemination, and user feedback of corrective information are relatively scarce. This makes it difficult for researchers to systematically study and validate relevant methods.

7 Conclusion

This study systematically reviews existing false information detection methods through the lens of information propagation on social networks. We begin by categorizing information dissemination into homogeneous and heterogeneous propagation. Based on this categorization, we introduce a comprehensive taxonomy that encompasses existing detection methods. We then provide a detailed overview and summary of related methods within these two categories, covering formal problem formulations, popular datasets, and detailed approaches, with an emphasis on the utilization of propagation dynamics and social context. Furthermore, we identify several promising avenues for future research, including the development of standardized evaluation frameworks, innovative strategies to mitigate the impact of false information, and the exploration of diverse information carriers.

Acknowledgements

This work is supported by the RGC GRF grant (No. 14217322) and the Tencent WeChat Rhino-Bird Focused Research Program.

References

- [1] Jonathan Bakdash, Char Sample, Monica Rankin, Murat Kantarcioglu, Jennifer Holmes, Sue Kase, Erin Zaroukian, and Boleslaw Szymanski. The future of deception: Machine-generated and manipulated images, video, and audio? In *2018 International Workshop on Social Sensing (SocialSens)*, pages 2–2. IEEE, 2018.
- [2] Fabiana Battista, Tiziana Lanciano, and Antonietta Curci. A survey on the criteria used to judge (fake) news in italian population. *Brain and Behavior*, 15(2):e70315, 2025.
- [3] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [4] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *AAAI*, pages 549–556, 2020.
- [5] Ya-Ting Chang, Zhibo Hu, Xiaoyu Li, Shuiqiao Yang, Jiaojiao Jiang, and Nan Sun. Dihan: A novel dynamic hierarchical graph attention network for fake news detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 197–206, 2024.
- [6] You Chang and Mart CM de Jong. A novel method to jointly estimate transmission rate and decay rate parameters in environmental transmission models. *Epidemics*, 42:100672, 2023.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
- [8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
- [9] Chaoqun Cui and Caiyan Jia. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *AAAI*, pages 73–81, 2024.
- [10] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862, 2020.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [12] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2051–2055, 2021.

- [13] Don Fallis. A functional analysis of disinformation. *IConference 2014 Proceedings*, 2014.
- [14] Shuzhi Gong, Richard O Sinnott, Jianzhong Qi, and Cecile Paris. Fake news detection through graph-based neural networks: A survey. *arXiv preprint arXiv:2307.12639*, 2023.
- [15] Shuzhi Gong, Richard Sinnott, Jianzhong Qi, and Cecile Paris. Unseen fake news detection through casual debiasing. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 981–985, 2025.
- [16] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36, 2020.
- [17] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [19] Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. Rumor detection on social media with event augmentations. In *SIGIR*, pages 2020–2024, 2021.
- [20] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113, 2024.
- [21] Fiza Gulzar Hussain, Muhammad Wasim, Seemab Hameed, Abdur Rehman, Muhammad Nabeel Asim, and Andreas Dengel. Fake news detection landscape: Datasets, data modalities, ai approaches, their challenges, and future perspectives. *IEEE Access*, 2025.
- [22] Wei Jiang, Tong Chen, Xinyi Gao, Wentao Zhang, Lizhen Cui, and Hongzhi Yin. Epidemiology-informed network for robust rumor detection. In *Proceedings of the ACM on Web Conference 2025*, pages 3618–3627, 2025.
- [23] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754, 2022.
- [24] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. Detection and analysis of 2016 us presidential election related rumors on twitter. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRIMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pages 14–24. Springer, 2017.
- [25] Zhezhou Kang, Yanan Cao, Yanmin Shang, Tao Liang, Hengzhu Tang, and Lingling Tong. Fake news detection with heterogenous deep graph convolutional network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 408–420. Springer, 2021.
- [26] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. In *2021 grace hopper celebration India (GHCI)*, pages 1–5. IEEE, 2021.
- [27] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8783–8790, 2020.
- [28] Junghoon Kim, Junmo Lee, Yeonjun In, Kanghoon Yoon, and Chanyoung Park. Revisiting fake news detection: Towards temporality-aware evaluation by leveraging engagement earliness. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 559–567, 2025.
- [29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [30] Srijan Kumar and Neil Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.
- [31] Lazarus Kwao, Yang Yang, Jie Zou, and Jing Ma. A survey of approaches to early rumor detection on microblogging platforms: Computational and socio-psychological insights. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e70001, 2025.
- [32] Batool Lakzaei, Mostafa Haghir Chehreghani, and Alireza Bagheri. Disinformation detection using graph neural networks: a survey. *Artificial Intelligence Review*, 57(3):52, 2024.
- [33] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer,

- Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
- [34] Angsheng Li and Yicheng Pan. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339, 2016.
 - [35] Xianghua Li, Jiao Qiao, Shu Yin, Lianwei Wu, Chao Gao, Zhen Wang, and Xuelong Li. A survey of multimodal fake news detection: A cross-modal interaction perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.
 - [36] Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Guang Chen. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. *arXiv preprint arXiv:2204.08143*, 2022.
 - [37] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. Mining significant microblogs for misinformation identification: an attention-based approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–20, 2018.
 - [38] Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*, 2024.
 - [39] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1867–1870, 2015.
 - [40] Guanghui Ma, Chunming Hu, Ling Ge, Junfan Chen, Hong Zhang, and Richong Zhang. Towards robust false information detection on social networks with contrastive learning. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1441–1450, 2022.
 - [41] Jiachen Ma, Jing Dai, Yong Liu, Meng Han, and Chunyu Ai. Contrastive learning for rumor detection via fitting beta mixture model. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4160–4164, 2023.
 - [42] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754, 2015.
 - [43] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.
 - [44] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics, 2017.
 - [45] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics, 2018.
 - [46] Alaa Safaa Mahdi and Narjis Mezaal Shati. A survey on fake news detection in social media using graph neural networks. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 16(2):23–41, 2024.
 - [47] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM web conference 2022*, pages 1148–1158, 2022.
 - [48] Mohamed Mostafa, Ahmad S Almogren, Muhammad Al-Qurishi, and Majed Alrubaian. Modality deep-learning frameworks for fake news detection on social networks: a systematic literature review. *ACM Computing Surveys*, 57(3):1–50, 2024.
 - [49] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.
 - [50] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347, 2021.
 - [51] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1732–1742, 2024.
 - [52] CV Negoita and DA Ralescu. On fuzzy optimization. *Kybernetes*, 6(3):193–195, 1977.
 - [53] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

- [54] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174, 2020.
- [55] Xingyu Peng, Junran Wu, Ruomei Liu, and Ke Xu. Rumor detection on social media with temporal propagation structure optimization. *arXiv preprint arXiv:2412.08316*, 2024.
- [56] Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, 139:110235, 2023.
- [57] Kai Shu, Amy Sliva, Suhan Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.
- [58] Kai Shu, Deepak Mahudeswaran, Suhan Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.
- [59] Yun-Zhu Song, Yi-Syuan Chen, Yi-Ting Chang, Shao-Yu Weng, and Hong-Han Shuai. Adversary-aware rumor detection. In *ACL/IJCNLP*, pages 1371–1382, 2021.
- [60] Mengzhu Sun, Xi Zhang, Jiaqi Zheng, and Guixiang Ma. Ddgc: Dual dynamic graph convolutional networks for rumor detection on social media. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4611–4619, 2022.
- [61] Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. Rumor detection on social media with graph adversarial contrastive learning. In *WWW*, pages 2789–2797, 2022.
- [62] Xiang Tao, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. Semantic evolvement enhanced graph autoencoder for rumor detection. In *WWW*, pages 4150–4159, 2024.
- [63] Gabrielle Taylor, Wenting Jiang, Xiao Qin, Ashish Gupta, et al. Misinformation detection: A survey of ai techniques and research opportunities. *Foundations and Trends® in Information Systems*, 8(2):66–147, 2024.
- [64] Jeffrey E Thomas. Statements of fact, statements of opinion, and the first amendment. *Calif. L. Rev.*, 74: 1001, 1986.
- [65] Lin Tian, Xiuzhen Jenny Zhang, and Jey Han Lau. Duck: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949, 2022.
- [66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [67] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088, 2010.
- [68] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380): 1146–1151, 2018.
- [69] Lanjun Wang, Zehao Wang, Le Wu, and An-An Liu. Bots shield fake news: Adversarial attack on user engagement based fake news detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2369–2378, 2024.
- [70] Rong-Hua Wang, Zhen Jin, Quan-Xing Liu, Johan van de Koppel, and David Alonso. A simple stochastic model with environmental transmission explains multi-year periodicity in outbreaks of avian flu. *PloS one*, 7 (2):e28873, 2012.
- [71] Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. Towards propagation uncertainty: Edge-enhanced bayesian graph convolutional networks for rumor detection. In *ACL/IJCNLP (1)*, pages 3845–3854, 2021.
- [72] Lingwei Wei, Dou Hu, Wei Zhou, Xin Wang, and Songlin Hu. Modeling the uncertainty of information propagation for rumor detection: A neuro-fuzzy approach. *IEEE Trans. Neural Networks Learn. Syst.*, 35 (2):2522–2533, 2024.
- [73] Jiaying Wu and Bryan Hooi. Decor: Degree-corrected social graph refinement for fake news detection. In *SIGKDD*, pages 2582–2593, 2023.
- [74] Haowei Xu, Chao Gao, Xianghua Li, and Zhen Wang. D2: Customizing two-stage graph neural networks

- for early rumor detection through cascade diffusion prediction. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 568–576, 2025.
- [75] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.
 - [76] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2253–2262, 2022.
 - [77] Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. *arXiv preprint arXiv:2406.02143*, 2024.
 - [78] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 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*, pages 1417–1423, 2021.
 - [79] Xuankai Yang, Yan Wang, Xiuzhen Zhang, Shoujin Wang, Huaxiong Wang, and Kwok Yan Lam. A macro-and micro-hierarchical transfer learning framework for cross-domain fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5297–5307, 2025.
 - [80] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 796–805. IEEE, 2019.
 - [81] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. Early detection of fake news by utilizing the credibility of news, publishers, and users based on weakly supervised learning. *arXiv preprint arXiv:2012.04233*, 2020.
 - [82] Yirong Zeng, Xiao Ding, Bibo Cai, Ting Liu, and Bing Qin. Exploring large language models for effective rumor detection on social media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2552, 2025.
 - [83] Guixian Zhang, Rongjiao Liang, Zhongyi Yu, and Shichao Zhang. Rumour detection on social media with long-tail strategy. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.
 - [84] Huaiwen Zhang, Xinxin Liu, Qing Yang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. T3rd: Test-time training for rumor detection on social media. In *Proceedings of the ACM Web Conference 2024*, pages 2407–2416, 2024.
 - [85] Jiajun Zhang, Zhixun Li, Qiang Liu, Shu Wu, Zilei Wang, and Liang Wang. Evolving to the future: Unseen event adaptive fake news detection on social media. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4273–4277, 2024.
 - [86] Kaiwei Zhang, Junchi Yu, Haichao Shi, Jian Liang, and Xiaoyu Zhang. Rumor detection with diverse counterfactual evidence. In *SIGKDD*, pages 3321–3331, 2023.
 - [87] Yusong Zhang, Kun Xie, Xingyi Zhang, Xiangyu Dong, and Sibao Wang. Rumor detection on social media with reinforcement learning-based key propagation graph generator. In *Proceedings of the ACM on Web Conference 2025*, pages 2742–2753, 2025.
 - [88] Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. Adversarial training: A survey. *arXiv preprint arXiv:2410.15042*, 2024.
 - [89] Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Jürgen Kurths. Propagation structure-aware graph transformer for robust and interpretable fake news detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4652–4663, 2024.
 - [90] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. arxiv 2016. *arXiv preprint arXiv:1610.07363*, 2016.
 - [91] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3): e0150989, 2016.
 - [92] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 109–123. Springer, 2017.

- [93] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *Acm Computing Surveys (Csur)*, 51(2):1–36, 2018.