Discover Computing

REVIEW

Open Access

Disinformation detection technology: a survey

Jinghui Peng[1*], Zitao Yang[2], Liwei Jia[3], Chenyang Shi[1] and Ping Hou[4]

*Correspondence:
Jinghui Peng
pjh20200828@163.com
[1]School of Artificial Intelligence, Anhui Polytechnic University, Wuhu 241000, Anhui, China
[2]China Industrial Control Systems Cyber Emergency Response Team, Beijing 100040, China
[3]Office of Educational Administration, Shenyang Aerospace University, Shenyang 150000, Liaoning, China
[4]The 71375th Unit of PLA, Harbin 150000, China

Abstract

In the era of Big Data, the proliferation of multi-source online information has made false information control a crucial element in sustaining a healthy digital ecosystem. The significant societal harm caused by misinformation has spurred academic interest in developing robust authenticity detection methods for online content. To date, three primary paradigms have emerged for authenticity detection: unimodal, multimodal, and external knowledge-based approaches. This work provides a detailed investigation into existing false information detection techniques, selecting representative studies to review the current research landscape. Furthermore, it organizes commonly used datasets and evaluation metrics in the field and identifies promising directions for future research in false information detection.

Keywords False information, True and false detection, Knowledge graph, Neural network, Attention mechanism

1 Introduction

It is well known that the malicious fabrication of disinformation that contradicts facts can lead to negative consequences. In recent years, with the development of communication technology and the popularity of intelligent mobile terminals, the scale of Internet information has shown explosive growth. The channels for people to post, transmit and obtain information are more convenient, and information can be disseminated in more ways and faster. Especially with the application of "We-media", the right to publish information has shifted from professionals to the general public, resulting in the spread of a series of unverified and false news. In addition, the malicious evaluation and attack by the Anti-fan and the Water Army also brought personal injury and other social governance problems, which then triggered and upgraded to network public opinion.

Network public opinion has become the main factor affecting social stability, and the generation of public opinion largely comes from false information. The detection and screening of false information on the Internet has become the main way to prevent and curb the occurrence and dissemination of public opinion. Internet false information mainly exists in News, We-media, comments and other media. The true and false detection of information is to automatically distinguish information through intelligent technology to realize the labeling of false information, so as to help Internet users and the public filter false information and avoid being misled and biased. In order to eliminate the influence of false information in network media, major social media in the world have successively implemented different rumor governance strategies. In China, the Center for Reporting illegal and Bad Information organized the "Internet joint rumor-refuting Platform"; Tencent launched the "Truth-checking" platform and the "WeChat rumor refuting" mini program; Weibo and the Ministry of Public Security to build a "National rumor-refuting platform"; Baidu created a comprehensive "Baidu refute rumors" open platform. In other countries, the Facebook security team used the same tactics to fight fake accounts to deal with the "Water Army"; Twitter closed more than 70 million accounts of the Water Army; Amazon has blocked more than 200 million suspected fake reviews in 2022, and in 2023 is experimenting with large models to help identify and combat fake reviews. Although the true and false detection of network information can be used as an effective means, the quality and effect of the detection are not always ideal, that is to say, in some cases there may be real information to be judged as false information, while missing false information to be judged as true.

In order to obtain the detection results of information and realize the discrimination of false information. In recent years, the detection methods such as neural network, multi-modal, knowledge graph have been proposed successively, providing new ideas for solving this problem. Multi-modal detection refers to the detection from a single text modal data to the fusion of multiple modal data such as text and pictures, including text and visual information in pictures. The neural network detection mainly refers to deep neural networks and large models, and uses deep models to mine the deep feature of information for measurement judgment to achieve detection. Knowledge graph detection refers to the recognition of truth and falsity through the known general or related domain highly reliable related knowledge.

This paper mainly introduces three common false information detection strategies, namely unimodal, multi-model and external knowledge, and analyzes the similarities and differences between different detection methods. To systematically guide this survey and address the core aspects of the field, we formulate the following research questions (RQs):

- *RQ1* What are the main technical paradigms in disinformation detection and how have they evolved?
- *RQ2* What are the comparative strengths and limitations of different detection approaches?
- *RQ3* What are the key challenges and promising future research directions?

The main contributions of this survey are fourfold: (1) We provide a systematic review of disinformation detection technologies following a structured methodology; (2) We present a comprehensive analysis of three main paradigms—unimodal, multimodal, and external knowledge-based approaches; (3) We identify and discuss critical challenges and limitations in current research; (4) We outline promising future research directions to guide further advancements in the field.

Section 2 introduces the survey methodology and related work. Section 3 provides a detailed analysis of the research status across three paradigms. Section 4 introduces commonly used datasets and evaluation metrics. Section 5 discusses the findings, limitations, and implications. Section 6 outlines future research directions. Section 7 concludes the paper.

2 Methodology and related work

2.1 Methodology

To ensure a comprehensive and systematic literature review, we adopted a methodology inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. While PRISMA is primarily designed for systematic reviews in medical and health-related fields, its core principles of transparency, reproducibility, and systematic reporting are highly valuable for surveys in computer science. However, we adapted the methodology to better suit the characteristics of the disinformation detection field, where rapid technological evolution and diverse publication venues necessitate a more flexible approach.

2.1.1 Data sources and search strategy

We conducted literature searches on major academic indexing databases, including IEEE Xplore, ACM Digital Library, Web of Science, and CNKI (for Chinese publications). The search focused on articles published between 2013 and 2025 to capture the most recent advancements. The key queries used a combination of keywords and their synonyms: ("fake news" OR "disinformation" OR "misinformation" OR "rumor detection") AND ("detection" OR "identification") AND ("deep learning" OR "neural network" OR "multimodal" OR "knowledge graph" OR "large language model").

2.1.2 Study selection and screening

The initial search returned over 1,200 articles. After removing duplicates, we screened the titles and abstracts of 850 articles for relevance. The inclusion criteria were: (1) the paper proposes a novel detection method or provides a significant review/survey; (2) the paper is published in a peer-reviewed journal or conference proceedings; (3) the full text is accessible. This screening resulted in 280 articles for full-text review. After a detailed assessment of these articles, 84 were finally selected for in-depth analysis and inclusion in this survey based on their technical contribution, innovation, and relevance to our research questions.

2.1.3 Data extraction and synthesis

Key information was extracted from the 84 selected articles, including the proposed method, technical paradigm, core innovation, datasets used, and main findings. Two annotators (the first and second authors) independently performed the data extraction and categorization. Disagreements were resolved through discussion until a consensus was reached. The Cohen's Kappa coefficient for inter-annotator agreement was 0.86, indicating a high level of consistency. The extracted studies were then clustered into three main technical paradigms: Unimodal Detection, Multimodal Detection, and External Knowledge-based Detection, with further sub-categorization within each paradigm to structure the review.

2.2 Definition and characteristics

Disinformation refers to untrue information, which may be intentionally or unintentionally distorted, altered or fabricated, thereby misleading or deceiving the masses and causing adverse and negative effects on society [1]. This information broadly includes false content such as rumors, hoaxes, and phishing attacks. False information has the basic characteristics of multiple output channels, fast propagation speed, wide spread range. At the same time, there are typical characteristics of information involving multiple fields, content diversity, multi-model form, dynamic evolution of process transmission, etc. The widespread existence of false information is not only an important content of network ecological governance, but also a hot topic concerned by information science and communication science.

From a societal perspective, the harms of disinformation are both profound and multidimensional. In the political realm, it can be weaponized to manipulate public opinion, interfere with electoral processes, and undermine political trust and social stability. Within the economic sphere, disinformation facilitates financial fraud and stock market manipulation, leading to substantial financial losses for individuals and markets alike. Public health efforts are also severely compromised, as evidenced by rumors concerning viruses and vaccines, which directly impede effective containment measures and endanger lives. Furthermore, disinformation erodes the foundational fabric of social trust, exacerbates intergroup polarization, and weakens social cohesion. Consequently, combating it has emerged as a pressing global challenge. "Deep faking," using artificial intelligence, is an advanced form of disinformation that can fake images, audio, video, and text to appear fake.

2.3 Core problem

With the popularization of communication technology and mobile terminals, artificial intelligence-driven fraud costs are lower and disinformation spreads faster, and manpower screening alone can no longer curb the spread of false information. With the attention of the academic and industrial circles to the problem of false information on the Internet, the automatic detection level of disinformation has been effectively improved. The initial paradigm for detection depended on manually engineered features and classical classifiers. Although interpretable, these methods were inherently constrained by their poor generalizability and strong reliance on domain expertise, making them ill-suited for the volatile nature of massive, multimodal data. This limitation motivated a pivotal shift to deep learning and multimodal approaches, which are capable of handling such complexity. Deep learning, in particular, represents a fundamental transition by leveraging end-to-end learning to autonomously derive rich semantic and non-semantic features directly from raw data. Illustratively, Graph Convolutional Networks model structural propagation patterns in social networks to flag anomalies, and Recurrent Neural Networks capture contextual relationships in text sequences.

2.4 Technical requirements

There are two kinds of inconsistency in disinformation: cross-modal inconsistency and content knowledge inconsistency. Cross-modal inconsistency refers to the inconsistency of information between different media or communication channels. For example, the same event may be described or presented differently in various media such as text, pictures, video, etc. This inconsistency may lead to confusion when the model processes information of different model, thus affecting the detection accuracy. The detection of cross-modal inconsistency necessitates the development of multimodal machine learning techniques. The significance of this approach lies in its capacity to transcend the limitations of unimodal analysis. By modeling the correlations and contradictions between different modalities—such as text, image, audio, and video—these techniques enable more precise identification of forged or misleading content. For instance, a model can be trained to determine whether the audio track in a news video is consistent with the speaker's lip movements, or to analyze if the content of an image substantiates its accompanying textual description.

Content knowledge inconsistency refers to the fact that the content of false information may contradict common sense or accepted facts. The inconsistency usually involves the credibility, authenticity and source reliability of information. Models trained purely on data are fundamentally limited in resolving discrepancies between content and real-world knowledge due to their inherent lack of contextual awareness. This critical shortfall is addressed by integrating external knowledge, a paradigm that enriches models with structured information from knowledge bases and semantic networks. This infusion of knowledge empowers models to conduct fact-based verification, such as evaluating the veracity of a statement attributed to a public figure by cross-referencing it with their consistently held, documented views.

In order to solve the above problems, the multi-modal fusion method can be used to integrate the information of different model, so that the model can better understand and compare the information of different channels. Moreover, knowledge base containing all kinds of facts and common sense can be constructed by using knowledge graph, semantic web and other technologies to provide reliable background information and reference basis for the model. Thus, the consistency detection of modal and content can be realized.

2.5 Detection process

Disinformation detection is more accurately characterized as an orchestrated process involving multiple stages than as a simple application of an individual model. In line with the preceding discussion, a comprehensive detection framework generally comprises four key phases, as shown in Fig. 1. This diagram outlines the complete workflow from initial data ingestion to the final deployment of results. The process begins with the concurrent collection of raw data and external knowledge. This is followed by

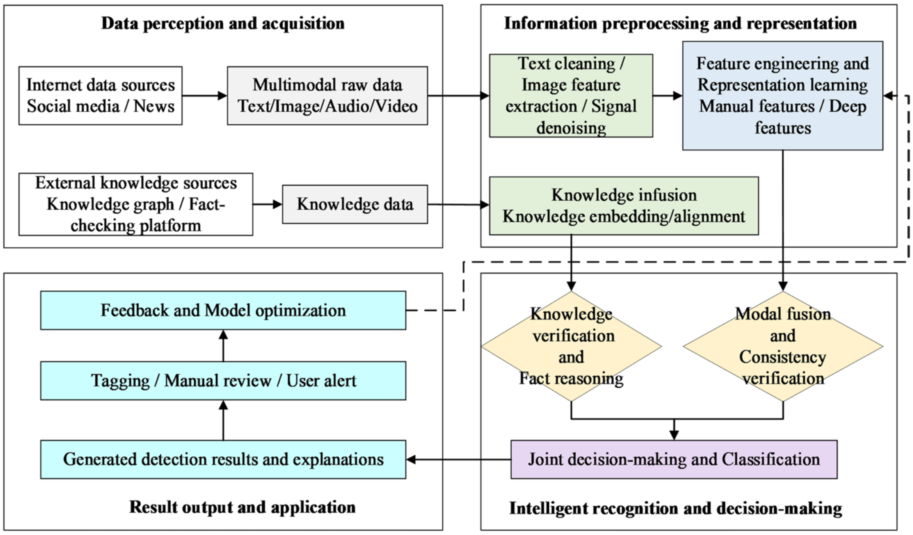
Data perception and acquisition
Internet data sources Social media / News
Multimodal raw data Text/Image/Audio/Video
External knowledge sources Knowledge graph / Fact-checking platform
Knowledge data
Information preprocessing and representation
Text cleaning / Image feature extraction / Signal denoising
Feature engineering and Representation learning Manual features / Deep features
Knowledge infusion Knowledge embedding/alignment
Knowledge verification and Fact reasoning
Modal fusion and Consistency verification
Joint decision-making and Classification
Intelligent recognition and decision-making
Feedback and Model optimization
Tagging / Manual review / User alert
Generated detection results and explanations
Result output and application

Fig. 1 Basic detection process

a representation phase where the inputs are preprocessed and transformed. The third and pivotal phase leverages both multimodal fusion and knowledge-based verification, whose outputs cooperatively support the classification decision. Finally, the generated results are applied in practical scenarios, while the feedback derived from these applications facilitates iterative model refinement to counter the dynamic nature of disinformation threats.

3 Research status

3.1 Unimodal detection

3.1.1 Multi-feature

Guo et al. [2] studied the detection of COVID-19 fake news by the two-classification method of support vector machine, which mainly includes the acquisition of data sources and the selection of data feature attributes, focusing on the comparison of the accuracy rate of different kernel functions in realizing non-separable non-linear in low-dimensional space and separable in high-dimensional space. Chang [3] combined text features with event features and proposed a transferable fusion detection model such as Fig. 2.

EANN in the model of Fig. 2 is a false information discrimination model based on GAN [4] method, which integrates three main components: feature extractor, false information detector and event discriminator. The feature extractor extracts text features through word embedding and text-CNN [5], and the false information detector accepts features as input to predict the truth and falsity of text messages. The event discriminator identifies the label of the event to which each text belongs based on the feature. Text-CNN represents the text as a stack of vectors through pre-trained word vectors, making the whole sentence into a matrix T. Convolve h terms from the i th word, get $t_i = \sigma\ (W_c{}^*T_{i:i+h-1})$. The fake news detector is a fully connected layer using Softmax. The feature obtained by G_f is sent to G_d for detection and the probability obtained is $P_\theta\ (m_i) = G_d\ (G_f\ (m_i;\ \theta_f);\ \theta_d)$. The event discriminator consists of two fully connected layers, represented as $G_e(R_f;\ \theta_e)$ to assign the correct data to the event.

3.1.2 Structural relationship

Zhang [6] proposed the FakeGFN model based on text feature and graph neural network, and introduced the “word-text-feature” graph into the prior feature to solve the global information loss problem of the current non-graph structure model. Lee et al. [7] leverage the Toulmin model of argumentation to construct a lexical network that captures the syntactic structure between claims and evidence. This network is subsequently transformed into a signed lexical network by incorporating semantic relationships. The

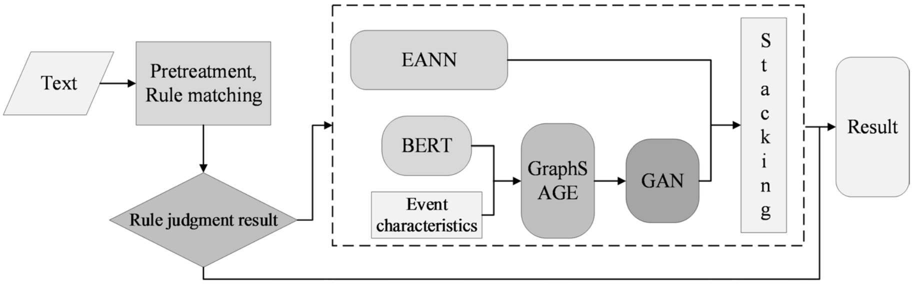

Fig. 2 A transportable fusion detection model

structural balance of the resulting network is then computed and used as a metric to quantify the coherence between claims and evidence. Wang [8] proposed SemSeq4FD model, a false information early detection model for sentences based on text structure. The model measures the structure between sentences: global semantic interaction, local adjacent context and global sequence feature. Firstly, the sentence is used as the node to encode the word embeddings are encoded as the initial sentence representation, and the fully connected complete graph is constructed. The graph convolutional neural network and self-attention mechanism are used to capture the global semantic relations between the sentences, so as to obtain the global sentence representation. Then, considering the contribution of the local context of adjacent sentences to the text expression, the text convolutional neural network is used to obtain the local sentence representation. Finally, the enhanced sentence representation is formed by concatenating the global and the local sentence representation, and the LSTM network is built to simulate the global reading order to merge the enhanced sentence representation and generate the final text representation for false information detection.

This structure captures both word order and history information when learning sentence representations using a Long short-term memory network (LSTM). Given a sentence S_i consisting of T_i words, input the word vectors from w_1 to w_T into the LSTM network in order. The model text as a semantic graph structure ζ=(v, ε), where nodes represent sentences in text and edges represent semantic relations between sentences. The graph convolution neural network (GCN) is used to aggregate the neighbor feature information to enhance the node representation. The initial feature matrix $\mathrm{X}^{(0)} \in R^{n^*m}$ and the adjacency matrix $\mathrm{A} \in R^{n^*n}$, and the graph convolution result of layer l is X^l=$tanh(AX^{(l-1)}W_s)$, where $W_s \in R^{n^*m}$ is the weight matrix to be learned. The dot-product self-attention method of formula (1) is used to enhance the sentence representation.

$$X_S = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Define the filter $w \in R^{k^*m}$, where k is the number of sentences in the sliding window and m is the sentence feature dimension. Moving a sliding window of size k sequentially from the first sentence to the last sentence using m' convolution filters gives a local sentence representation matrix $X_a \in R^{n^*m'}$. Splicing X_s and X_a horizontally gives the joint representation $\mathrm{S_e} \in R^{n^*2m'}$.

3.1.3 Neural network

Based on RNN, Chen et al. [9] proposed a deep memory model that can learn continuous hidden representations by capturing remote dependencies and context changes. Text was first transformed into TF-IDF vectors, and RNN networks are enhanced with self-attention mechanisms to obtain temporal potential representations by capturing long-term dependencies between sequences and selectively focusing on important correlations. The forgetting gate Z^f control in LSTM unit selectively forgets the c^{t-1} input of the previous state. Each cell learns to weigh the contribution of its input gate and modulation to the input modulator, as well as the weight of the memory released simultaneously. The attention module calculates the average value of the current input x_t as a TF-IDF feature based on attention Softmax a_t weighting. The model takes feature slice x_t as input, propagates x_t through the stacking layer of LSTM, and predicts the next position probability of a_{t+1} and class label y_t. Hu et al. [10] integrate message content and propagation pathways to construct a dual-directional graph convolutional network for enhanced fake news detection. By transforming disconnected user posts into a bidirectional propagation graph, the approach leverages network propagation topology. Contextual semantic features are extracted from source texts and their propagated counterparts using BERT embeddings, which are subsequently incorporated as node attributes into the propagation graph.

3.1.4 Transformers

Early approaches heavily relied on feature engineering. Word embeddings, such as Word2Vec and GloVe, were pivotal in representing textual semantics for models like SVM and Random Forests. These representations allowed classifiers to capture some semantic meaning, but were often shallow. The advent of Transformers marked a paradigm shift, enabling models like BERT to capture deep contextual relationships in text. Their self-attention mechanism provides a significant advantage over RNNs in modeling long-range dependencies, making them highly effective for fake news detection [11]. Subsequent research has explored fine-tuned transformers [12], sentence transformers for better semantic textual similarity [13], and document embeddings for holistic document representation [14]. More advanced architectures, such as those using disentangled attention and enhanced mask decoders (e.g., mDeBERTa) [12], have further pushed the state-of-the-art by enabling more precise modeling of content and context.

3.2 Multi-modal detection

3.2.1 Shallow fusion model

Li [15] preprocessed the text, visual and speech information and extracted the features. The three modal features, *t*, *v* and *a*, were fused and compared by concatenating Z_c=[t;v;a], Hadamard product Z_h=[t⊙v⊙a] and Kronecker product $Z_k = [t \otimes v \otimes a]$. BiLSTM network was used to extract text features, ResNet50 network to extract visual features, and Baidu API converted speech into text features. Wang et al. [16] proposed a knowledge-driven multi-modal graph convolution network (KMGCN) framework, as shown in Fig. 3. Keywords were extracted from the text, combined with knowledge

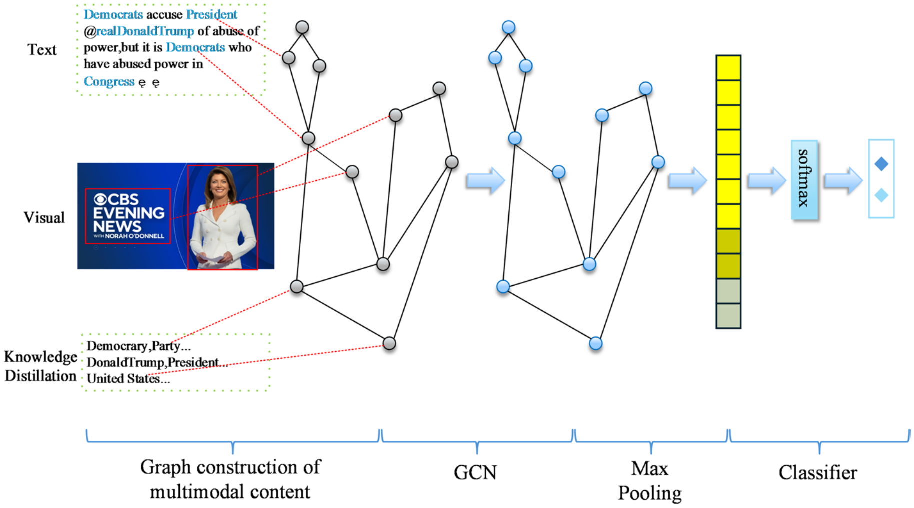
Text
Democrats accuse President @realDonaldTrump of abuse of power,but it is Democrats who have abused power in Congress ẹ ẹ
Visual
CBS EVENING NEWS
WITH NORAH O'DONNELL
Knowledge Distillation
Democrary,Party...
DonaldTrump,President...
United States...
softmax
Graph construction of multimodal content
GCN
Max Pooling
Classifier

Fig. 3 Framework of knowledge-driven multi-modal graph convolutional networks

distillation to complement the background knowledge of keywords, and directly connected with YOLOv3 detector from image detection semantics, which was modeled into an undirected graph according to the weight of calculated edge of point mutual information (PMI). The nodes of each graph are aggregated using two GCN layers and a global mean pooling.

3.2.2 Deep fusion model

Chen [17, 18] designed a deep learning multi-modal false news detection model that integrated knowledge graph and image description. This model mined the knowledge graph in the form of triples implied by news texts and obtained its distributed representation through knowledge graph embedding technology. At the same time, the deep description text corresponding to the image is mined, and the text is generated for distributed vector representation. Then Bert framework is used to integrate two distributed representations of knowledge graph and image description. Liang [19] uses the multilayer CNN model to fuse multi-modal information from the semantic and feature levels, and uses the corresponding classifier to obtain the probability distribution, and obtains the final prediction result by weighting. Tong et al. [20] propose a multi-modal and multi-domain fake news detection model(MMDFND), which integrates domain embeddings and attention mechanisms into a progressive hierarchical extraction network to enable domain-adaptive extraction of relevant knowledge.

Meng [21] simultaneously learns the interaction between multiple modes and the complex relationships within single modes from a global perspective, and proposes a multi-model fine-grained fusion false information detection model TMF based on Transformer-MF such as Fig. 4. The module is composed of a multi-head self-attention mechanism and a feed-forward network. The process begins with a masked multi-head attention layer applied to the input, which is then processed by an "Add & Norm" layer. A feed-forward network follows, succeeded by a second "Add & Norm" operation. This complete block is replicated N times to form the final structure. The model can learn the interaction between word and word, word and region, and region and region, and obtain global and local multi-modal representation. The feature extraction module is used to extract the representation of each word in the text and the representation of each region in the image to achieve the refinement of the information fusion granularity between the text and image.

3.2.3 Learning contrast

Li [22] proposed a multi-modal false information detection method based on contrastive learning pre-training and attention mechanism. Bidirectional encoder representation pre-training model (BERT) was used to extract language features from modal data of test text, and the residual network (ResNet) was used to extract image features from modal data of test image. Contrast learning is used to map different modal feature vectors to the same feature space by minimizing contrast loss function among different modal data for feature alignment and potential relationship learning. Attention mechanism is used to realize high-level interaction of different modal features, and feature fusion is used to complete model construction to detect false information.

Xu et al. [23] put forward similarity and relative similarity strategies for detection. By combining positive and negative pairs' information through comparative learning, rich

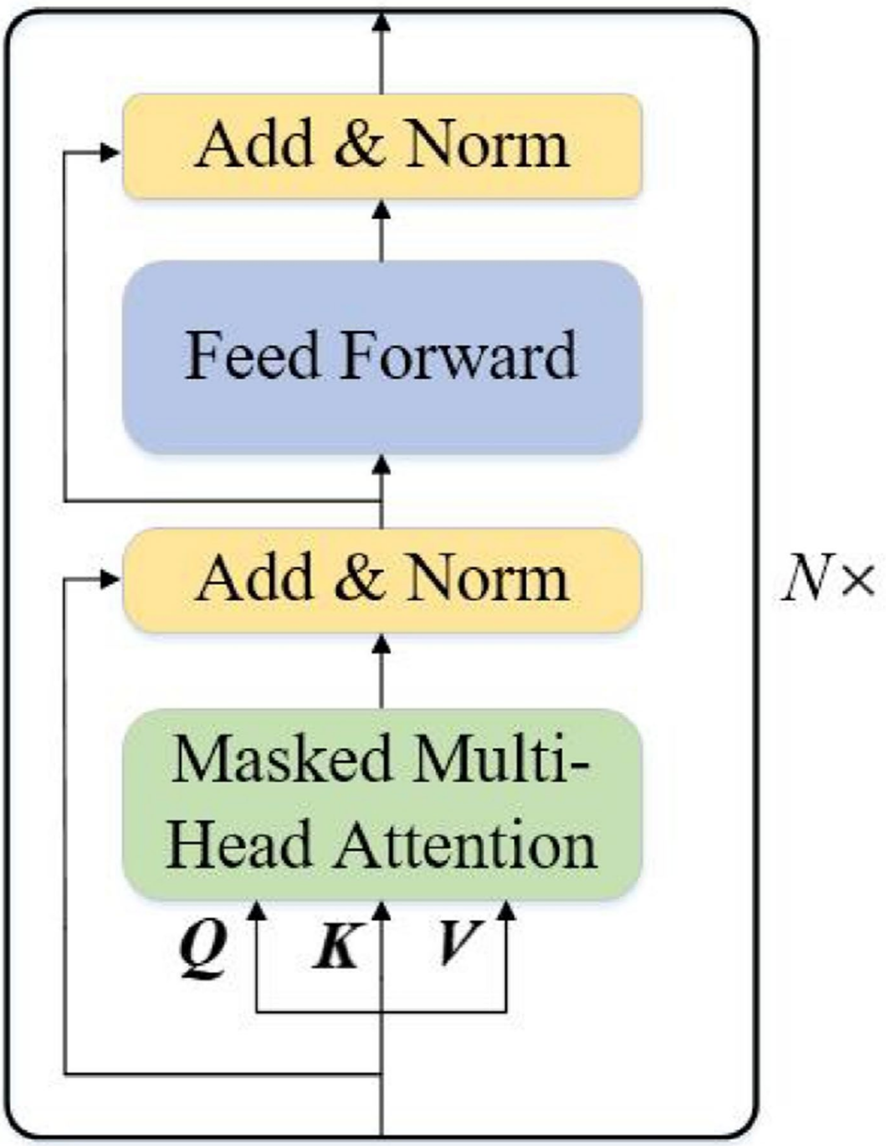

Fig. 4 Illustration of the transformer-MF module

discriminant representations can be learned and the intrinsic features of sentences can be captured. Liao [24] proposed a cross-document false information detection model (CAL) based on contrast graph learning, which combines the contrast learning module for expanding the differential representation of facts and false information in vector space and the heterogeneous graph module for infusing the semantic features of public opinion environment. Graph neural network (GNN) [25] encodes heterogeneous graphs, and message transfer on the graphs is realized through aggregation and merging, as shown in formula (2).

$$\boldsymbol{u}_i^n = f_s\left(\boldsymbol{h}_i^n\right) + \sum\nolimits_{r\in R} \frac{i}{|N_i^r|} \sum\nolimits_{j\in N_i^r} f_r\left(\boldsymbol{h}_j^n\right) \tag{2}$$

3.2.4 Semantic enhancement

Wang [26] took the single text feature BFID model as the baseline to study the false information detection method that combines emotional features and image features to enhance semantics. Firstly, BERT model is used for text representation, and then text matrix is input into BiLSTM + Attention module. Then the emotion analysis model is obtained through BERT + BiLSTM model training, and the output of the intermediate layer vector is extracted as the supplementary feature of the integration of emotion factors, and it is spliced with the text vector matrix. In order to be consistent with the text vector dimension, image features are obtained using ResNet18 network and then spliced. Fang et al. [27] identified fake news content through news semantic context awareness (NSEP). First, the NSEP discriminant module will use depth with time constraints between the semantics of the environment is divided into the macro semantic E^{mac}={ v_1^k, v_2^k,..., v_j^k} and microcosmic semantic environment E^{mic}={ v_1^k, v_2^k,..., v_r^k}. Then the graph convolutional network is used to perceive the content feature G in the macro semantic environment. Then the micro-semantic detection module guided by multi-head attention and sparse attention is used to capture the content feature F in the micro-semantic environment. The last two vectors are cascaded with the original detector and then fed into the MLP and Softmax as shown in formula (3) to obtain the final prediction.

$$\widehat{y} = softmax\,(MLP(\,F \oplus \,G \oplus \,Detector_{text}\,)) \tag{3}$$

3.3 External knowledge detection

3.3.1 Knowledge graph

Fang [28] proposed a multi-granularity bidirectional LSTM text analysis network that extracted four types of entities, and combined the model iteration with time series to measure the relationship between entities, so as to build a dynamic graph network for detecting fake comments. Gao et al. [29] propose a knowledge-augmented vision-and-language model to enhance multimodal fake news detection. The model integrates information from large-scale open knowledge graphs to improve its ability to assess the veracity of news content. Unlike previous approaches that rely on separate models for textual and visual feature extraction, this work introduces a unified framework capable of concurrently processing both modalities.

Sun [30] built a multi-modal anti-common sense false information detection model (KDIN) and a dynamic propagation structure false information detection model (DDGCN) based on the external knowledge of knowledge graphs. The former built a neural network through the double inconsistence of cross-modal and content knowledge, while the latter built a double dynamic graph structure by using the spatial and temporal structure of information propagation. Pi et al. [31] proposed a detection method based on knowledge graph representation learning. Firstly, entity and relation representation were obtained through PN-KG2REC algorithm. Then, the entity and relation representation in the triplet to be detected are concatenated to obtain the triplet representation, and the vertex entity e starts to perform L random walks, and a walk sequence of length $2L + 1$ can be obtained, $\boldsymbol{S}$ ={$s^{(1)}$, $s^{(2)}$, ..., $s^{(2\text{ L }+\text{ }1)}$}, the entity-relation sequence $\boldsymbol{S}'$is shown in Eq. (4). Finally, the vector representations of triples are classified.

$$\boldsymbol{S}' = (e_{\phi\,\left(s^{(1)}\right)}, r_{\Phi\,\left(s^{(2)}\right)}\,, \cdots, e_{\phi\,\left(s^{(2l-1)}\right)}, r_{\Phi\,\left(s^{(2l)}\right)}\,, \cdots, e_{\phi\,\left(s^{(2L+1)}\right)}) \tag{4}$$

Sun et al. [32] proposed a new dual dynamic graph convolution algorithm network (DDGCN), which used the network in a unified framework to model the dynamics of message propagation and the background knowledge of knowledge graph in different time stages to integrate structural information and time units. Which consists of three main parts: (1) Building module of dynamic graph, constructing time rumor propagation graph and time event-entity-concept tripartite knowledge graph respectively. (2) A graph convolution network module composed of bistatic GCN units and temporal fusion units is adopted to obtain the structural semantic features of events, and to fuse and disseminate information and knowledge information at each time stage. (3) Classification module, which summarizes the final propagation, knowledge and text information, and generates classification labels.

3.3.2 Meta-prompt

Huang et al. [33] proposed a meta-prompt learning (MAP) framework. By constructing templates, the detection task is transformed into a pre-training task to tap into the potential of pre-trained language models. To address the issue that randomly initialized templates affect the mining performance, the advantage of meta-learning in rapid parameter training is borrowed to learn the optimal initialization parameters. Meta-learning and prompt learning are combined for detection, and meta-tasks are constructed to obtain initialization parameters suitable for different domains, achieving the establishment of a prompt model language instrument for classification in low-resource noisy scenarios. For the former, a multi-domain meta-task construction method is used to learn domain-invariant meta-knowledge. For the latter, a prototype language instrument is used to summarize category information, and an anti-noise prototype strategy is designed to reduce the impact of noisy data. In the meta-training stage, domain-invariant meta-tasks are constructed using source data to train the prompt model. In the meta-testing stage, a small amount of target domain data is used to fine-tune the meta-trained model and verify the results.

3.3.3 Pre-trained Mmodel

Mu et al. [34] defined information X={(R, P, U)} detection as a regression task, developed an open source domain adaptive pre-trained language model, and used post and user-level information to evaluate the performance of BERT-Weibo-Rumor and several supervised classifiers. Using CN-DBpedia [35] as the external knowledge graph, H1, H2, H3, and H4 are projected into a dense vector of the same dimension, and the final combination of inputs obtained by fusion represents H5. H5 through the fully connected layer, the rumor prevalence prediction is obtained using the standard mean square error (MSE) loss function. All model parameters were fine-tuned, including two different BERT encoders for the rumor content (R) and the user profile description (P).

Jiang et al. [36] proposed a similar sensing multi-modal prompt learning framework (SAMPLE). To combine prompt learning into multi-modal fake news detection, three analyzers with soft language are used. In addition, a similarity sensing fusion method is introduced, which adaptively fuses the intensity of multi-modal representations to mitigate noise injection from uncorrelated cross-modal features. In order to identify the most relevant images corresponding to the text of a given news article, the pre-trained CLIP model [37] is used to encode the text and image representations respectively. In order to reduce the coarse feature dimension provided by the encoder and eliminate redundant information, a single projection head P_{txt} and P_{img} are used to process text and image features. Each projection head has two sets of fully connected layers and is batch normalized, rectified linear unit (ReLU) activation function and dropout layer. The cosine similarity between the two feature variables C_sim is measured to correct the intensity of the fusion features $\mathrm{f}_{\mathrm{fused}}$, and only the image with the highest similarity to the text representation is retained.

3.3.4 Large Language model

Hu et al. [38] designed an Adaptive Rationale-Guided network (ARG) for fake news detection, in which a Small Language Model (SLM) selectively distills insights for news analysis from rationales generated by a Large Language Model (LLM). A rationale-free variant of ARG, termed ARG-D, is derived through distillation to accommodate cost-sensitive scenarios that preclude queries to the LLM. Alqadi et al. [39] introduced a multi-phase transfer learning framework that leverages pre-trained LLMs, specifically RoBERTa, for fake news detection in data-scarce scenarios. Diverging from prior studies that primarily relied on standard fine-tuning, their approach systematically compares word embedding techniques such as Word2Vec and one-hot encoding, combined with a refined fine-tuning process to enhance both model performance and interpretability. Jin et al. [40] proposed the CAPE-FND framework in fake news detection (FND), which employs Context-Aware Prompt Engineering. This framework mitigates LLM hallucinations through veracity-oriented contextual constraints, background information, and analogical reasoning. It further refines LLM predictions via adaptive guided prompt optimization and enhances prompt effectiveness through an adaptive iterative process using a stochastic search-guided algorithm.

To enhance model capacity and generalization, researchers have incorporated architectures like the Mixture of Experts (MoE) [41]. These models dynamically route inputs to specialized "expert" networks, allowing for scalable and efficient handling of diverse disinformation patterns. Variants include the Adaptive Mixture of Transformers [42] and Language-based Mixture of Transformers [43], which tailor the mixture mechanism for specific modal or linguistic tasks. Furthermore, ensemble methods that combine multiple models have proven highly effective. For instance, some studies integrate predictions from content-based models with features from social network propagation embeddings [44] or user interaction graphs [45], creating robust systems that are less vulnerable to the limitations of any single model.

3.4 Comparison of several methods

False information detection models can be classified into classification algorithms, unimodal, multi-modal, knowledge graph and large models based on their types. Different types can be further subdivided into various specific methods. This section introduces the contents, advantages and disadvantages of these methods. The specific information is shown in Table 1.

According to the insights from Table 1, the adaptability of early detection paradigms—those utilizing hand-crafted features and traditional machine learning (e.g., SVM, Random Forest)—is severely limited. Their performance is inextricably tied to the completeness of feature engineering, making them ill-suited for automating the detection

Table 1 Comparison of different methods

Types	Methods	Factors	Pros and cons
Classification algorithm	SVM [46]	Manual features [47]	It is effective for high-dimensional data, but lacks comprehensiveness and flexibility
	RF [48]	Feature importance	High robustness and anti-overfitting; However, the model has poor interpretability.
	DT [49]	Node splitting rule	Intuitive and easy to explain; But it is prone to overfitting.
Unimodal	Feature extraction [50]	User [51]、linguistics[52]、emotion[53]、location-time [54]	The feature selection is diverse, but the risk of redundancy is high.
	Kernel function [55]	Weight, score	It is capable of handling nonlinear problems, but the selection of kernel functions and parameter tuning are difficult.
	Graph model [56]	Random variables and hyperparameters	The ability of relationship modeling is strong, but the training complexity is high.
	RNN [57]	Hide status	The sequence modeling is effective, but the problem of vanishing gradients is significant.
	Attention mechanism [58]	Weights allocation	It has a strong ability to focus on key information, but the computational overhead is large.
	GAN [59]	Generator and discriminator	It has strong data generation ability, but the training is unstable.
Multi-modal	attRNN [60]	Attention-weighted fusion	Cross-modal time series modeling is excellent, but the number of parameters is large.
	EANN [61]	Event adversarial learning	Event-independent noise suppression, but the adversarial training is difficult.
	Concatenation of models [62]	Feature cascading	The implementation is simple and fast, but the modal interaction capability is weak.
	Decompose the bilinear pool model [63]	Modal interaction matrix	Fine-grained feature fusion, but with high computational complexity.
KG	Trans method [63]	Entity relationship embedding	Simple and efficient, but difficult to model complex relationships.
	GNN [64]	Neighborhood aggregation	The structural information is strongly retained, but the risk of over-smoothing is high.
	Rule-based reasoning [65]	Logical constraints	It has strong interpretability, but the rule coverage is limited.
	Knowledge distillation [66]	Teacher-student structure	The model is lightweight, but knowledge loss is inevitable.
LLM	Transformer [67]	Self-attention mechanism	The global dependency modeling is strong, but the video memory consumption is huge.
	MoE [68]	Expert routing approach	It has strong scalability, but the risk of unbalanced training is high.

of dynamically evolving disinformation. In contrast, deep learning models (e.g., RNN, attention mechanisms) achieve greater generalization through end-to-end learning. Notably, the Transformer architecture excels at modeling contextual relationships, thereby enabling it to better discern semantic contradictions in novel disinformation campaigns. However, a critical limitation persists: these deep learning models are inherently passive. Their knowledge is frozen at the point of training and lacks the capacity for active integration of external, up-to-date knowledge. This stands in stark contrast to the dynamic nature of disinformation itself.

4 Data and evaluation indicator

Based on the comparison and analysis of the existing work, this section introduces the commonly used datasets and common evaluation indicators for false information detection.

4.1 Dataset

FND dataset [36]: Accessing the PolitiFact and GossipCop datasets of the FakeNewsNet project [69], 1056 and 22,140 news data were obtained, respectively, by eliminating redundancy and saving news with images. The results were 198 (96 true and 102 false) and 6,805 (1,187 true and 4,928 false) news items, respectively.

Weibo 21 dataset [70]: Collect the multi-domain Chinese dataset of Sina and Weibo from 2014 to 2021. Nine areas (science, Military, Education, disaster, Politics, Health, Finance, Entertainment and social), which contained 4,488 fake news articles and 4,640 real news articles.

Chinese dataset [71]: The collection includes selected news clips from 2010 to 2018, with 16,349 true and false data (8913 true and 7436 false).

PHEME dataset [72]: Tweets of related events are obtained primarily through Twitter. It contains five events: the Charlie Hebdo shooting, the Germanwings plane crash, the Ferguson riots, the Ottawa shooting, and Sydney College, with a total of 5802 messages (3830 true and 1972 false).

Prz Dataset [73]: A fake news detection corpus collected from over 200 websites, with a total of 894 pieces of information (302 true and 592 false).

4.2 Evaluation indicator

In this paper, no matter how the intermediate process is operated, the final output of regression classification is 0 and 1. Therefore, Accuracy, Precision, Recall and F1 are usually used as evaluation indexes in experimental tests.

4.3 Test benchmark

The commonly used benchmark models including: (1) TextCNN [74] which utilizes the CNN for the task of false news classification; (2) TextRCNN [75] applied to the text classification task; (3) FastText [76] a relatively rapid method for word vectors and text classification; (4) LSTM [77] used for the false news classification task; (5) SVM-RBF [78] a classification detection model based on SVM and combined with the RBF kernel; (6) RvNN [79] a model based on the tree-structured RNN; (7) VAE-GCN [80] a detection model based on GCN with graph convolutional encoding and decoding; (8) Bi-GCN [81] a detection model based on GCN; (9) PPC [82] a detection model combining RNN and CNN; (10) HAGNN [83] a detection model based on GNNs; (11) GCNFEM [84] a detection model using GCNs.

5 Discussion and limitations

This survey has systematically reviewed the landscape of disinformation detection technologies, addressing the research questions posed in the introduction. For RQ1, we identified and detailed three main technical paradigms—Unimodal, Multimodal, and External Knowledge-based detection—and traced their evolution from feature-based classifiers to sophisticated neural architectures and knowledge-enhanced approaches.

The field has progressed from relying on manually crafted features to leveraging deep learning for automatic feature extraction, and more recently toward integrating external knowledge to overcome the limitations of purely data-driven models.

For RQ2, our comparative analysis reveals distinctive strengths and limitations across paradigms. Unimodal methods (Sect. 3.1) offer computational efficiency but struggle with the complexity of real-world misinformation. Multimodal approaches (Sect. 3.2) enhance robustness through cross-modal verification but face challenges in effective fusion and are vulnerable to sophisticated cross-modal forgeries. External knowledge methods (Sect. 3.3), particularly those using KGs and LLMs, enable fact-based verification and better generalization but depend heavily on the quality, coverage, and timeliness of external knowledge sources.

5.1 Key challenges and limitations

Despite significant progress, our analysis reveals several persistent challenges that limit current approaches:

Generalization and Domain Adaptation Models often exhibit performance degradation when applied to new domains, topics, or cultural contexts. The dynamic nature of disinformation tactics creates a continuous "concept drift" problem that challenges static models.

Explainability and Interpretability The "black-box" nature of many deep learning models hinders trust and practical deployment, particularly in scenarios where understanding the rationale behind detection decisions is crucial.

Data Scarcity and Imbalance High-quality, large-scale annotated datasets remain scarce, especially for low-resource languages and emerging events. This leads to models biased toward well-represented domains and classes.

Real-time Detection Performance The computational complexity of sophisticated multimodal and knowledge-based models often impedes their application in real-time detection scenarios where rapid response is critical.

Adversarial Robustness Malicious actors continuously develop adversarial examples designed to evade detection systems, creating an ongoing arms race between attackers and defenders.

Knowledge Currency and Integration For KG and LLM-based methods, the latency between real-world events and their incorporation into knowledge bases remains a significant challenge, as does the effective integration of dynamic, streaming knowledge.

Ethical and Privacy Concerns Personalized detection approaches raise serious privacy issues regarding user data collection and analysis. There is also potential for bias in automated detection systems that must be carefully addressed.

5.2 Interplay of paradigms and research gaps

A key insight from this survey is that no single paradigm provides a complete solution. The most promising approaches appear to be hybrid systems that intelligently combine the strengths of multiple paradigms. For instance, using LLMs for semantic understanding and rationale generation, grounded by structured knowledge from KGs, and validated against multimodal evidence could create more robust and explainable systems. However, significant research gaps remain in effectively integrating these components, particularly in developing frameworks that can adaptively weigh different types of evidence and provide transparent reasoning processes.

6 Future research direction

6.1 Cross-language detection

The Internet era often leaves room for the generation of false information due to language differences, and its scope of influence and potential harm will increase exponentially. Studying the expression of information in different language environments, transmission channels and audience characteristics, and analyzes the unique rules of cross-language information transmission. To explore the influence of cultural differences, language barriers and information interpretation bias on the spread of cross-language false information. Cross-language false information detection can curb the spread of false information from the source of information and eliminate the gap between language differences. Through the learning and training of the deep learning model, the deep features of multi-lingual data can be automatically extracted and corresponding feature vectors can be generated. These feature vectors can effectively capture the intrinsic structure and semantic information of data, and provide a richer and more comprehensive representation for subsequent tasks. By selecting suitable feature fusion method and similarity calculation method, the correlation and consistency between different languages can be evaluated effectively.

Current research faces several critical challenges: the scarcity of annotated data for low-resource languages leads to inadequate model generalization; significant differences in cultural contexts and expression habits across languages increase the difficulty of semantic alignment; and existing cross-lingual representation learning methods remain insufficient in fine-grained sentiment and stance recognition. To address these challenges, future studies may focus on the following technical pathways: developing transfer learning frameworks based on cross-lingual pre-trained models to enhance model adaptability to low-resource languages through adversarial training and knowledge distillation; constructing graph neural network models that integrate linguistic knowledge and semantic representations to achieve deep semantic alignment of cross-lingual information; and designing multi-task learning mechanisms that jointly perform language identification, sentiment analysis, and credibility assessment to improve overall system performance.

6.2 Personalized detection

The traditional detection of false information usually relies on the text content, the reputation of the publisher, and the mode of dissemination of the information. However, these methods rarely take into account individual users' preferences, trust tendencies, and information processing habits. In order to solve this problem, personalized feature extraction becomes a key step. By analyzing the user's personal information, behavior track and interactive data, it can extract the indicators that reflect their personalized characteristics, such as trust network, information processing habits, emotional tendencies and so on. Deep learning techniques can be used to identify individual users' language patterns, emotional colors, and information processing habits to more accurately judge the authenticity of information.

The core of personalized detection lies in establishing an associative model between user cognitive characteristics and information credibility. This direction faces several key technical challenges: effectively extracting individual features without infringing on user privacy, balancing personalized recommendations with the mitigation of information cocoon effects, and achieving real-time modeling of users' dynamic interests. Promising technical pathways may focus on the following aspects: developing federated learning-based user modeling frameworks to extract users' trust tendencies and behavioral patterns while preserving data privacy; leveraging graph neural networks to construct user-information bipartite graphs and identify information processing characteristics of similar user groups through community detection algorithms; and designing temporal dynamic models to capture the evolution of users' cognitive habits. Particular attention should be given to enhancing model interpretability to ensure transparency and logical consistency in personalized detection outcomes.

6.3 Cross-modal deep feature consistency detection

When false information spreads between different model, its influence and potential harm will be further expanded. For the data of different modes, such as text, image, audio and video, the corresponding deep learning model needs to be used for feature extraction. These models can automatically learn and extract deep features of multimodal data, including semantic, visual and auditory features. By feature extraction, the original data can be transformed into high-dimensional feature vectors. Appropriate feature fusion methods are adopted, such as series, concatenation, weighted summation, etc. These methods can combine features of different modes to form a unified feature representation. By comparing the feature similarity and correlation between different model, it can judge whether with consistent false information characteristics.

Cross-modal detection faces the core issues of bridging the semantic gap and achieving feature alignment across modalities. Key technical challenges include significant discrepancies in feature distributions among different modal data, information redundancy and noise interference during multimodal fusion, and the difficulty in capturing fine-grained cross-modal semantic relationships. Future research should prioritize breakthroughs in the following areas: developing hierarchical feature extraction networks based on cross-modal attention mechanisms to achieve fine-grained alignment of text, image, audio, and other modalities; constructing multimodal contrastive learning frameworks that enhance the model's ability to perceive consistent features through positive and negative sample pairs; and designing multimodal knowledge graphs to explicitly model semantic relationships across modalities. Particular attention should be devoted to developing lightweight detection solutions to meet the demands of real-time application scenarios.

6.4 Domain expertise knowledge-guided detection

False information detection based on domain expertise is a combination of domain expertise and machine learning technology to improve the accuracy and reliability of false information detection. First, it is necessary to conduct in-depth understanding and research in related fields and sort out the professional knowledge in this field. This includes the understanding of the concepts, principles, laws and other aspects of the field, as well as the common problems and misunderstandings in the field. By sorting out the professional knowledge in the field, it can better understand the manifestations and dissemination characteristics of false information, and provide guidance and support for the subsequent feature extraction and model construction, especially the research on the guidance mechanism of integrating professional knowledge into the model.

This research direction seeks to deeply integrate domain knowledge into detection models to improve accuracy in specialized scenarios. Key technical challenges include the difficulty in formalizing and quantifying domain knowledge, the unclear mechanisms for integrating expert knowledge with data-driven models, and insufficient domain adaptation capabilities. To enable effective knowledge-guided detection, potential approaches involve constructing domain-specific knowledge graphs and leveraging graph neural networks to learn structured knowledge representations; developing knowledge-enhanced pre-training frameworks that incorporate domain rules as constraints during model training; and designing hybrid reasoning mechanisms that combine symbolic reasoning with neural network predictions. A critical focus lies in addressing the synergy between knowledge updates and model iteration, thereby establishing a sustainable and evolvable knowledge-guided detection system.

7 Conclusion

This paper has provided a comprehensive examination of disinformation detection technologies, systematically addressing the defined research questions. We traced the field's evolution from early feature-based methods to sophisticated multimodal fusion and the emerging use of external knowledge through knowledge graphs and large language models (RQ1). Our comparative analysis revealed the distinctive advantages and limitations of each paradigm, highlighting that robustness often emerges from combining data-driven learning with structured, factual knowledge (RQ2). Furthermore, we identified critical challenges—including generalization limitations, explainability deficits, real-time performance constraints, and adversarial vulnerabilities—that currently constrain state-of-the-art systems (RQ3).

The survey demonstrates that while substantial progress has been made, disinformation detection remains a dynamic and challenging research area. The proliferation of generative AI and sophisticated cross-modal forgeries demands continuous innovation and adaptation. Future advancements will likely depend on developing more adaptive, explainable, and efficient hybrid models that leverage the complementary strengths of multimodal data analysis and external knowledge integration. The research directions outlined in Sect. 6, including cross-lingual detection, privacy-preserving personalized modeling, deep cross-modal consistency verification, and domain-knowledge integration, provide a roadmap toward next-generation detection systems.

Ultimately, effectively combating disinformation requires not only technological innovations but also multidisciplinary collaboration across computer science, social science, psychology, and policy-making. As detection technologies advance, so too do the techniques for creating and disseminating false information, necessitating a continuous cycle of research, development, and adaptation. This survey contributes to this effort by providing a comprehensive overview of the current state of the art and a structured framework for guiding future research in this critically important field.

Acknowledgements
The authors wish to thank Fei Fan, and Junjie Sun for their comments and help.

Author contributions

P. is responsible for the overall design and writing of the paper; Y. and J. is responsible for the data and architecture of the paper; S. wrote and drew the paper; H. Helped to revise the manuscript.

Funding

This work was supported by the Key Laboratory of the Ministry of Industry and Information Technology with Grant Nos. KY20250113.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

We consent to the publication of anonymized data/results collected from our participation in this study. We understand that all personally identifiable information will be removed to protect our privacy. Additionally, we grant permission for the use of non-identifiable data in presentations, reports, journals, and other academic publications.

Informed consent

We voluntarily agree to participate in this study titled "Disinformation Detection Technology: A Survey". We confirm that we have been informed about the nature, purpose, procedures, potential risks and benefits of the study, and that any questions we have raised were answered to our satisfaction.

Competing interests

The authors declare no competing interests.

Received: 25 June 2025 / Accepted: 11 December 2025

Published online: 24 December 2025

References

1. Wang J, Wang YC, Huang MJ. False in formation in social networks: definition, detection and control. J Comput Sci. 2019;48(8):263–77.
2. Guo WQ, Li B. False news detection in the background of COVID-19 based on SVM. J Foshan Univ Sci Technol (Natural Sci Edition). 2021;39(06):19–26.
3. Chang CS. Research and implementation of a social network rumor detection system based on multi-feature fusion. Beijing University of Posts and Telecommunications; 2022.
4. Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: an overview. IEEE Signal Process Mag. 2018;35(1):53–65.
5. Kim Y. Convolutional neural networks for sentence classification. ArXiv preprint arXiv:1408.5882, 2014.
6. Zhang HR. Research on fake information detection methods based on text feature and graph neural network. Central University of Finance and Economics; 2022.
7. Lee K, Ram S. Explainable deep learning for false information identification: an argumentation theory approach. Inf Syst Res. 2024;35(2):890–907.
8. Wang YH. Research on fake news detection based on text structure. Taiyuan University of Technology; 2022.
9. Chen T, Wu L, Li X et al. Call attention to rumors: deep attention based recurrent neural networks for early rumor detection. ArXiv Preprint arXiv: 1704.05973, 2017.
10. Hu J, Yang M, Tang B. Integrating message content and propagation path for enhanced false information detection using bidirectional graph convolutional neural networks. Appl Sci. 2025;15(7):3457.
11. Truică CO, Apostol ES. Misrobærta: transformers versus misinformation. Mathematics. 2022;10(4):569.
12. Cotelin MD, Truică CO, Apostol ES. NetGuardAI at EXIST2025: sexism detection using mDeBERTa. Working Notes of CLEF; 2025.
13. Truică CO, Apostol ES, Paschke A. Awakened at CheckThat! 2022: Fake news detection using BiLSTM and sentence transformer. CEUR workshop proc. 2022, 3180: 749–757.
14. Truică CO, Apostol ES. It's all in the embedding! Fake news detection using document embeddings. Mathematics. 2023;11(3):508.
15. Li GH. Research on false content detection model of network media. Harbin Engineering University; 2024.
16. Wang Y, Qian S, Hu J et al. Fake news detection via knowledge-driven multimodal graph convolutional networks. Proceedings of the 2020 international conference on multimedia retrieval. 2020: 540–547.
17. Chen KY, Xu F, Wang MW. The fake news detection based on knowledge graph and image description. J Jiangxi Normal Univ (Natural Sci edition). 2021;45(04):398–402.
18. Chen KY. Research on multi-modal fake news detection based on knowledge graph. Jiangxi Normal University; 2022.
19. Liang Y, Turdi T, Eschal A. Multi-modal false information detection via multi-layer CNN-based feature fusion and multi-classifier hybrid prediction. Comput Eng Sci. 2023;45(06):1087–96.
20. Tong Y, Lu W, Zhao Z et al. MMDFND: Multi-modal multi-domain fake news detection. Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 1178–1186.
21. Meng J. Research on false information detection based on multimodal fusion. Taiyuan University of Technology; 2022.
22. Li ZY, Li J. Contrastive learning-based multimodal attention networks for false information detection method. Chin Sci Pap. 2023;18(11):1192–7.

23. Xu L, Xie H, Wang FL, et al. Contrastive sentence representation learning with adaptive false negative cancellation. Inf Fusion. 2024;102:102065.
24. Liao JZ, Zhao HW, Liao XT et al. Contrastive graph learning for cross-document misinformation detection. Comput Sci, 2023: 1–9.
25. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[C]. Annual Conference on Neural Information Processing Systems (NeurIPS), 2017: 1024–1034.
26. Wang H, Gong LJ, Zhou ZJ, et al. Detecting mis/dis-information from social media with semantic enhancement. Data Anal Knowl Discovery. 2023;7(2):48–60.
27. Jadhav P, Shukla RK. Deep learning analysis for revealing fake news using linguistic complexity and semantic signatures. Int J Intell Syst Appl Eng. 2024;12(12s):458–65.
28. Fang YL. Research on detection method of false comments based on knowledge graph. Shandong Normal University; 2019.
29. Gao X, Wang X, Chen Z, et al. Knowledge enhanced vision and language model for multi-modal fake news detection. IEEE Trans Multimedia. 2024;26:8312–22.
30. Sun MZ. Research and implementation of false information detection technology based on knowledge graph. Beijing University of Posts and Telecommunications; 2022.
31. Pi DC, Wu ZY, Cao JJ. Early rumor detection method based on knowledge graph representation learning. Acta Electron. 2023;51(02):385–95.
32. Sun M, Zhang X, Zheng J et al. Ddgcn: Dual dynamic graph convolutional networks for rumor detection on social media. Proceedings of the AAAI conference on artificial intelligence. 2022, 36(4): 4611–4619.
33. Huang Y, Gao M, Wang J, et al. Meta-prompt based learning for low-resource false information detection. Information Processing & Management. 2023;60(3):103279.
34. Mu Y, Niu P, Bontcheva K, et al. Predicting and analyzing the popularity of false rumors in Weibo. Expert Syst Appl. 2024;243:122791.
35. Xu B, Xu Y, Liang J et al. CN-DBpedia: A never-ending Chinese knowledge extraction systemInternational Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer International Publishing, 2017: 428–438.
36. Jiang Y, Yu X, Wang Y, et al. Similarity-aware multimodal prompt learning for fake news detection. Inf Sci. 2023;647:119446.
37. Radford A, Kim JW, Hallacy C et al. Learning transferable visual models from natural language supervisionInternational conference on machine learning. PMLR, 2021: 8748–8763.
38. Hu B, Sheng Q, Cao J et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection. Proceedings of the AAAI conference on artificial intelligence. 2024, 38(20): 22105–22113.
39. Alqadi BS, Alsuhibany SA, Yousafzai SN, et al. Transfer learning driven fake news detection and classification using large language models. Sci Rep. 2025;15(1):28490.
40. Jin W, Gao Y, Tao T, et al. Veracity-oriented context-aware large language models–based prompting optimization for fake news detection. Int J Intell Syst. 2025;2025(1):5920142.
41. Petrescu A, Truică CO, Apostol ES. Language-based mixture of Transformers for EXIST2024. Working Notes of CLEF; 2024.
42. Petrescu A, Apostol ES, Truică CO. Awakened at EXIST2025: adaptive mixture of Transformers. Working Notes of CLEF; 2025.
43. Petrescu A, Truică CO, Apostol ES. Language-Based Mixture of Transformers for Sexism Identification in Social Networks. International Conference of the Cross-Language Evaluation Forum for European Languages. Cham: Springer Nature Switzerland, 2025: 142–155.
44. Truică CO, Apostol ES, Marogel M, et al. GETAE: graph information enhanced deep neural network ensemble architecture for fake news detection. Expert Syst Appl. 2025;275:126984.
45. Truică CO, Apostol ES, Karras P. DANES: deep neural network ensemble architecture for social and textual context-aware fake news detection. Knowl Based Syst. 2024;294:111715.
46. Pérez-Rosas V, Kleinberg B, Lefevre A et al. Automatic detection of fake news. ArXiv Preprint arXiv:1708.07104, 2017.
47. Wang SH, Zhang YD, Yang M, et al. Unilateral sensorineural hearing loss identification based on double-density dual-tree complex wavelet transform and multinomial logistic regression. Integr Comput Aided Eng. 2019;26(4):411–26.
48. More S, Idrissi M, Mahmoud H, et al. Enhanced intrusion detection systems performance with UNSW-NB15 data analysis. Algorithms. 2024;17(2):64.
49. Teo TW, Chua HN, Jasser MB et al. Integrating large Language models and machine learning for fake news detection.2024 20th IEEE international colloquium on signal processing & its applications (CSPA). IEEE, 2024: 102–7.
50. Farhangian F, Cruz RMO, Cavalcanti GDC. Fake news detection: taxonomy and comparative study. Inf Fusion. 2024;103:102140.
51. Castillo C, Mendoza M, Poblete B. Information credibility on twitter.Proceedings of the 20th international conference on World wide web. 2011: 675–684.
52. Kwon S, Cha M, Jung K et al. Prominent features of rumor propagation in online social media.2013 IEEE 13th international conference on data mining. IEEE, 2013: 1103–1108.
53. Wu K, Yang S, Zhu KQ. False rumors detection on sina weibo by propagation structures.2015 IEEE 31st international conference on data engineering. IEEE, 2015: 651–662.
54. Ma J, Gao W, Wong KF. Detect rumors in microblog posts using propagation structure via kernel learning[C]. Association for Computational Linguistics; 2017.
55. Kokate S, Chettyi MSR. Fraudulent event detection in credit card data operations using SVM-RBF kernel function machine learning classification technique. Intell Decis Technol. 2025. https://doi.org/10.1177/18724981241305118.
56. Yang S, Shu K, Wang S et al. Unsupervised fake news detection on social media: A generative approach.Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 5644–5651.
57. Ma J, Gao W, Mitra P et al. Detecting rumors from microblogs with recurrent neural networks. (2016).Proceedings of the 25th international joint conference on artificial intelligence (IJCAI 2016). 2016: 3818–3824.
58. Qi P, Cao J, Yang T et al. Exploiting multi-domain visual information for fake news detection.2019 IEEE international conference on data mining (ICDM). IEEE, 2019: 518–527.
59. Ma J, Gao W, Wong KF. Detect rumors on twitter by promoting information campaigns with generative adversarial learning.The world wide web conference. 2019: 3049–3055.

60. Jin ZW, Cao J, Guo H et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs.Proceedings of the 25th ACM international conference on Multimedia. 2017: 795–816.
61. Wang Y, Ma F, Jin ZW et al. Eann: Event adversarial neural networks for multi-modal fake news detection.Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining. 2018: 849–857.
62. Singhal S, Shah RR, Chakraborty T et al. Spotfake: A multi-modal framework for fake news detection.2019 IEEE fifth international conference on multimedia big data (BigMM). IEEE, 2019: 39–47.
63. Kumari R, Ekbal A. Amfb: attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Syst Appl. 2021;184:115412.
64. Ma G, Hu C, Ge L et al. Towards robust false information detection on social networks with contrastive learning.Proceedings of the 31st ACM international conference on information & knowledge management. 2022: 1441–1450.
65. Nguyen QA, Pham TT, Trinh VG et al. Detecting Misleading Information with LLMs and Explainable ASP.International Conference on Agents and Artificial Intelligence ICAART 2025. 2025.
66. Wang C, Meng L, Xia Z, et al. Cross-Domain deepfake detection based on latent domain knowledge distillation. IEEE Signal Processing Letters; 2025.
67. Huang T, Xu Z, Yu P et al. A Hybrid Transformer Model for Fake News Detection: Leveraging Bayesian Optimization and Bidirectional Recurrent Unit. arxiv preprint arxiv:2502.09097, 2025.
68. Ma Y, Yu Z, Lin X et al. BIG-MoE: bypass isolated gating MoE for generalized multimodal face Anti-Spoofing. arxiv preprint arxiv:2412.18065, 2024.
69. Shu K, Mahudeswaran D, Wang S, et al. Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data. 2020;8(3):171–88.
70. Nan Q, Cao J, Zhu Y et al. MDFEND: Multi-domain fake news detection.Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 3343–3347.
71. Zhu Y, Sheng Q, Cao J et al. Generalizing to the future: Mitigating entity bias in fake news detection. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 2120–2125.
72. Zubiaga A, Liakata M, Procter R. Learning reporting dynamics during breaking news for rumour detection in social media. ArXiv Preprint arXiv:1610.07363, 2016.
73. Liu Y, Wu YFB. Fned: a deep network for fake news early detection on social media. ACM Trans Inform Syst (TOIS). 2020;38(3):1–33.
74. Yoon Kim. Convolutional neural networks for sentence classification [EB/OL]. [2025-06-02]. https://arxiv.org/abs/1408.5882
75. Lai Siwei X, Liheng L, Kang et al. Recurrent convolutional neural networks for text classification [EB/OL]. [2025-06-02]. https://ieeexplore.Ieee.org/document/8852406
76. Armand J, Edouard G, Piotr B et al. Bag of tricks for efficient text classification [EB/OL]. [2025-06-02]. https://arxiv.org/abs/1607. 01759.
77. Bahad P, Saxena P, Kamal R. Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Comput Sci. 2019;165:74–82.
78. Yang F, Yang L et al. Yu Xiaohui,. Automatic detection of rumor on sina weibo. Proceedings of the ACM SIGKDD workshop on mining data semantics. 2012: 1–7.
79. Jing M, Wei G, Wong KF. Rumor detection on Twitter with tree structured recursive neural networks [C]. Beijing, China: Association for Computational Linguistics; 2018. pp. 8–13.
80. Lin Hongbin Z, Xi F, Xianghua, A Graph Convolutional Encoder and Decoder Model for Rumor Detection. 2020 IEEE the 7th International Conference on Data Science and Advanced Analytics. IEEE, 2020, 67 (3): 300–306.
81. Tian B, Xi X, Tingyang X et al. Rumor detection on social media with bi-directional graph convolutional networks. Proceedings of the AAAI conference on artificial intelligence. 2020, 549–556.
82. Liu Yang W. Yifang. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. Proceedings of the AAAI conference on artificial intelligence. 2018, 32 (1): 1–8.
83. Xu Shouzhi L, Xiaodi M. Rumor detection on social media using hierarchically aggregated feature via graph neural networks. Appl Intell. 2023;53:3136–49. https://doi.org/10.1007/s10489-02203592-3.
84. Thota NR, Xiaoyan S. Dai Jun. (2023). Early Rumor Detection in Social Media Based on Graph Convolutional Networks. 2023 International Conference on Computing, Networking and Communications (ICNC), 516–522.

Publisher's note