Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.059536

ARTICLE

From Detection to Explanation: Integrating Temporal and Spatial Features for Rumor Detection and Explaining Results Using LLMs

Nanjiang Zhong*, Xinchen Jiang and Yuan Yao

College of Information and Network Security, Zhejiang Police College, Hangzhou, 310053, China

*Corresponding Author: Nanjiang Zhong. Email: zhongnanjiang@zjjcxy.cn

Received: 10 October 2024; Accepted: 18 December 2024; Published: 06 March 2025

ABSTRACT: The proliferation of rumors on social media has caused serious harm to society. Although previous research has attempted to use deep learning methods for rumor detection, they did not simultaneously consider the two key features of temporal and spatial domains. More importantly, these methods struggle to automatically generate convincing explanations for the detection results, which is crucial for preventing the further spread of rumors. To address these limitations, this paper proposes a novel method that integrates both temporal and spatial features while leveraging Large Language Models (LLMs) to automatically generate explanations for the detection results. Our method constructs a dynamic graph model to represent the evolving, tree-like propagation structure of rumors across different time periods. Spatial features are extracted using a Graph Convolutional Network, which captures the interactions and relationships between entities within the rumor network. Temporal features are extracted using a Recurrent Neural Network, which accounts for the dynamics of rumor spread over time. To automatically generate explanations, we utilize Llama-3-8B, a large language model, to provide clear and contextually relevant rationales for the detected rumors. We evaluate our method on two real-world datasets and demonstrate that it outperforms current state-of-the-art techniques, achieving superior detection accuracy while also offering the added capability of automatically generating interpretable and convincing explanations. Our results highlight the effectiveness of combining temporal and spatial features, along with LLMs, for improving rumor detection and understanding.

KEYWORDS: Rumor detection; graph convolutional neural networks; recurrent neural networks; large language models

1 Introduction

The development of social media has greatly improved the communication efficiency, but it has also inadvertently contributed to the rapid spread of rumors and caused serious social impact. During key events such as the US presidential election [1] and the COVID-19 epidemic [2], rumors flooded social media, causing social chaos. For example, Islam et al. [2] found that during the COVID-19 pandemic, 82% of the news information between 31 December 2019, and 05 April 2020, consisted of rumors. These rumors included false information about government control measures, incorrect treatments, and the origins of the disease, causing severe negative impacts on both individuals and the government. This highlights the urgent need for effective strategies to detect rumors and block the spread of rumors through reasonable explanations.

The initial approaches to rumor detection primarily relied on analyzing the text of the rumors. Early methods evolved from traditional machine learning [3–5] to deep learning [6–9] and achieved some degree

of success. However, since rumor texts are often brief and can be easily disguised as normal content, detection methods based solely on text analysis have certain limitations in terms of accuracy.

As a result, more and more methods [10] have started to focus on rumor detection based on propagation structures, as these are more difficult to falsify. These methods generally fall into two categories when constructing propagation structure models: temporal-domain-based methods and spatial-domain-based methods. Temporal-domain-based methods [11–13] build a temporal propagation structure based on the chronological order of reposts or comments, as shown in Fig. 1b. On the other hand, spatial-domain-based methods [14–16] construct a tree-like propagation network based on the interaction relationships between reposts or comments, as shown in Fig. 1c.

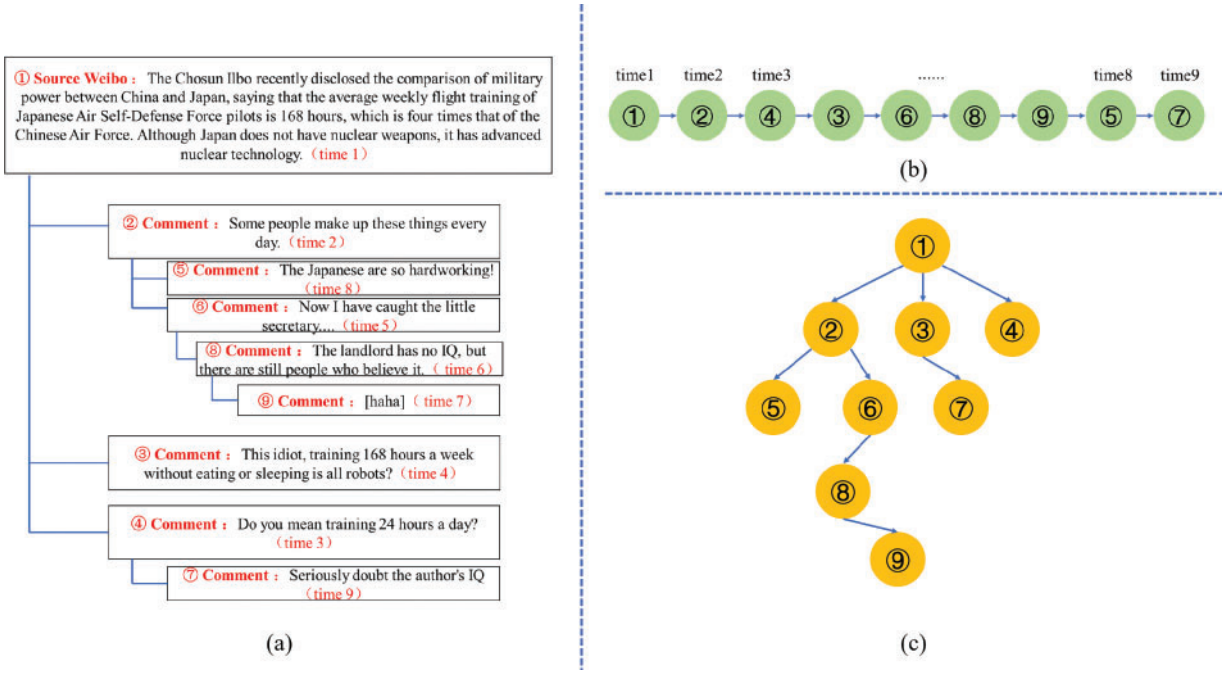

Figure 1: Traditional rumor propagation modeling approaches. (a) Hierarchical structure of rumor comments: ① represents the original tweet of the event, and the lower-level text is a comment on the upper-level text. For example, ② is a comment on ①, and ⑤ is a comment on ②. The "time n" at the end of the text indicates the time when the comment was posted, with n increasing from smallest to largest to represent the chronological order of comment postings. (b) Modeling method based on temporal features: The sequence structure is constructed based on the chronological order of comment postings. (c) Modeling method based on spatial features: A propagation network is constructed based on the relationships between comments, and rumors are detected by extracting the spatial features of the propagation network

However, the existing methods have the following main issues: 1) They fail to effectively integrate the temporal and spatial features of rumor propagation. Some methods merely combine temporal and spatial features by concatenating them, without capturing the dynamic changes in the propagation structure over time. 2) They lack reasonable explanations for the detection results. Deep learning methods are notably deficient in interpretability, yet providing a clear and reasonable explanation for the detection results is crucial for effectively curbing the spread of rumors.

To address the aforementioned issues, this paper proposes a rumor detection method that integrates both temporal and spatial features, and provides reasonable explanations for detection results using a

LLM. For rumor detection, we first construct a dynamic propagation graph model by generating multiple propagation graphs based on repost and comment relationships at different time points. Then, we employ a Graph Convolutional Network (GCN) to extract spatial features from the propagation graph at each discrete time point. Finally, we capture the temporal features of the propagation network by feeding the features from different time points into a Gated Recurrent Unit (GRU) model. In this process, we are able to effectively capture the temporal evolution of the rumor spatial propagation network.

We selected LLM as the interpreter, using carefully crafted prompts to input the detection results, rumor text, and propagation information, thereby generating a reasonable explanation for the detection results. In this process, we divided the propagation information into a propagation chain, allowing the LLM to perform step-by-step reasoning in the order of comment timestamps, enabling more effective analysis and interpretation of the rumor propagation process.

To evaluate the effectiveness of our method, we conducted experiments on the Weibo and PHEME datasets. Even without applying any data filtering, the experimental results still showed that our method significantly outperforms existing approaches. In terms of explaining the detection results, multiple case studies demonstrated that our method can provide clear and convincing explanations for the detection outcomes.

Our main contributions are as follows:

- We propose a novel rumor detection method that constructs a dynamic propagation graph model and utilizes GCN and GRU models to extract rumor features, significantly improving detection accuracy.
- We introduce a LLM to provide convincing explanations for the rumor detection results, which is crucial for effectively curbing the spread of rumors.
- We conducted extensive experiments comparing various state-of-the-art rumor detection methods, and the results indicate that our method performs better.

2 Related Work

2.1 Temporal-Domain-Based Methods

As social media spreads globally, the challenge of detecting misinformation and rumors has intensified. Temporal-domain-based methods, particularly those using time series analysis, have gained significant attention for identifying rumors by analyzing the propagation patterns of information.

For instance, one approach [17] uses Recurrent Neural Networks (RNNs) to capture the erratic, burst-like patterns typical of rumor spread, contrasting with the steady progression of truthful information. This innovation lies in applying RNNs to model the specific temporal dynamics of rumor propagation, offering an adaptive detection mechanism.

Another method [18] combines time series data with user behavior analysis, examining reposting rates and the speed of information spread. Rumors tend to have faster, more volatile lifespans in early propagation stages, which serves as a key indicator. The novelty of this approach is integrating both temporal data and user behavior to improve detection robustness.

Additionally, integrating social network features with time series analysis has proven effective for early-stage rumor detection [19]. This approach analyzes both temporal behavior and social interactions, offering a more nuanced understanding of rumor propagation. Its contribution lies in combining these two dimensions to enhance early detection accuracy.

Furthermore, multi-modal models, such as the Capture, Score, and Integrate (CSI) framework [7], combine content, social context, and temporal features for comprehensive detection. By incorporating emotional tone, user behavior, and timing, these models provide more robust mechanisms for rumor

identification. The novelty here is the multi-dimensional approach, leveraging a wide range of features for enhanced detection.

Overall, temporal-domain-based methods offer a powerful tool for identifying rumors by analyzing the dynamic patterns of information propagation within social networks.

2.2 Spatial-Domain-Based Methods

Spatial-domain-based rumor detection methods focus on analyzing the structural features of information dissemination networks. These methods use graph theory and network topology to examine the relationships between nodes (users) and identify potential rumors.

One approach [20] constructs a graph of information propagation and applies graph embedding techniques to capture the network's topological features. By analyzing node connectivity and subgraph structures, this method can distinguish between rumors and truthful information. Research shows that rumors often form densely connected subgraphs, which are key indicators of misinformation. The novelty lies in focusing on these subgraphs to track rumor spread.

Another approach [21] uses social network graph models to analyze node interactions, particularly the role of influential nodes (e.g., opinion leaders) in rumor spread. Studies show that rumors tend to spread rapidly among these key nodes, with concentrated dissemination paths. The novelty of this method is its focus on influential node dynamics, which helps predict and control rumor spread more effectively.

Additionally, Graph Convolutional Networks (GCNs) have been used [14] to analyze spatial features of social networks. GCNs examine the relationships between neighboring nodes to detect rumors by analyzing information flow and interaction patterns. This method is particularly effective in detecting subtle signals of rumor spread in complex networks. The novelty of GCNs lies in their ability to handle large, intricate networks, making them highly effective for real-time rumor detection.

Overall, spatial-domain-based methods use graph theory and network structures to identify rumors by analyzing how information propagates and how key nodes interact, offering a powerful tool for detecting misinformation.

2.3 Large Language Models

Large language models (LLMs), such as GPT-3 and its successors, have revolutionized applications in natural language processing and conversational AI. However, their ability to generate highly realistic, human-like text poses significant challenges for misinformation detection [22]. As LLMs produce content that closely mimics human writing, traditional detection methods, which rely on linguistic features like structure and syntax, often fail to distinguish between machine-generated and human-written misinformation [23,24].

Several studies have explored using LLMs in misinformation detection. Yang et al. [25] utilize GPT-3.5 to extract entities and build relational graphs, enhancing the identification of false information. While effective, it still faces challenges in accuracy and scope. Hu et al. [26] highlight a key limitation: fine-tuned, task-specific models for fake news detection outperform general-purpose LLMs, suggesting that LLMs' versatility may hinder their efficiency in detecting misinformation.

The growing sophistication of LLMs has spurred interest in hybrid models that combine their capabilities with specialized detection mechanisms. These models aim to improve detection at scale, but LLMs still struggle with capturing nuanced context, addressing adversarial manipulation, and adapting to new domains without fine-tuning. As LLMs continue to evolve, further research is needed to enhance their effectiveness in misinformation detection.

In conclusion, while LLMs offer advances in natural language generation, their use in misinformation detection remains complex. Current efforts leverage LLMs alongside other models, but overcoming their limitations will require further innovation.

3 Problem Definition

Let $C = \{c_1, c_2, \ldots, c_m\}$ be the rumor detection dataset, where c_i is the i-th event and m is the number of events. A event is defined as $c_i = \{r_0^i, r_1^i, \ldots, r_{n_i}^i, y_i\}$, where $n_i + 1$ represents the total number of microblogs involved in event c_i, r_0^i is the source microblog, r_j^i represents the jth responsive microblog in event c_i, such as comments and reposts. each forwarded or commented post r_j^i contains three dimensions of information, which are the text content w_j^i, the posting time t_j^i, and the parent node p_j^i. Here, $y_i \in Y$, $Y = \{R, N\}$ represents the ground-truth of event c_i, where R represents rumor and N represents normal event. Given the above dataset, our goal is to model the propagation process and extract the propagation features of events and learn a classifier:

$$f: c_i \rightarrow y_i \tag{1}$$

For a clearer expression, we will omit the event number i in the subsequent formula description.

4 Methodology

4.1 Model Overview

The overall structure of the method in this paper is shown in Fig. 2. The proposed rumor detection model consists of five main components, dynamic graph construction, encoding text contents, spatial feature extraction, temporal feature extraction and classification with MLP.

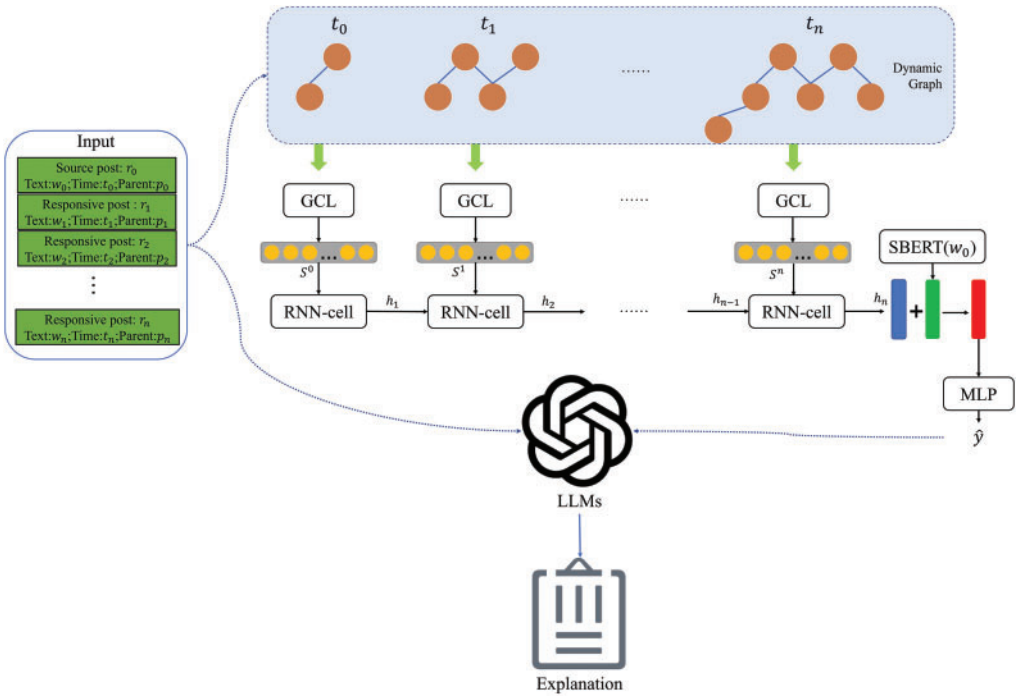

Figure 2: The framework of our approach

First, we construct a dynamic propagation structure graph of events, and then use graph convolutional neural networks to extract spatial features of propagation structure. Then we use the spatial domain features as input and use the recurrent neural network to extract the temporal domain features. At this time, we get a representation of the fusion of spatial domain and temporal domain features. Then, we integrate the text features of the source microblog to enhance the final feature representation. Finally, MLP is used to classify microblog events.

4.2 Dynamic Graph Construction

In this subsection, we construct a dynamic propagation structure graph of events. Unlike ordinary propagation structure graph models, which only build a single complete propagation structure graph, we construct multiple propagation graphs based on time series.

The choice of dynamic graph construction is motivated by the need to capture the evolving nature of rumor propagation over time. Rumors spread through social networks are not static; they evolve as new comments, reposts, and interactions occur. By modeling this progression dynamically, we can better understand how rumors unfold in real-time, allowing us to capture more nuanced patterns of rumor spread.

As shown in Fig. 3, the nodes represent the source microblog and its comments in the event, and the edges between the nodes represent their comment relationships. For example, r_1 commented on r_0 at time t_2, so there is a directed edge between them. As time progresses, the propagation structure graph evolves dynamically, with the network's structural features becoming increasingly rich. For instance, at time t_0, the propagation graph contains only the source tweet as a single node, whereas by time t_n, the graph incorporates the complete set of propagation structure features.

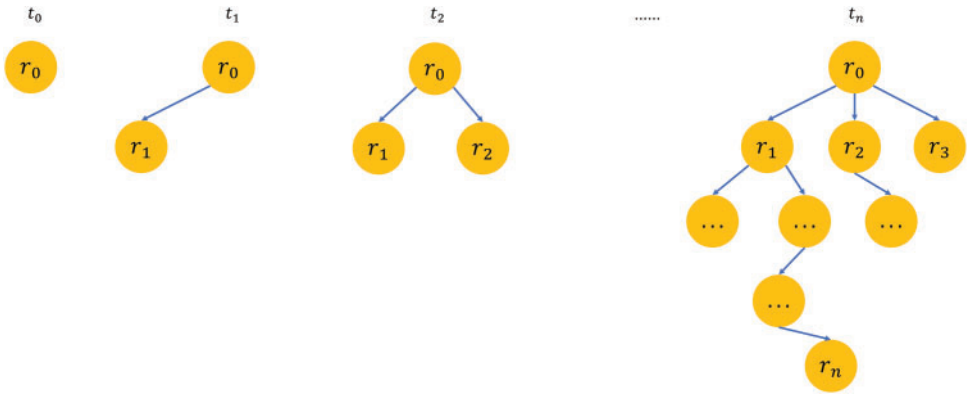

Figure 3: Schematic diagram of dynamic graph construction

In this view, we model each event as a set of directed graphs $G = \{G_0, G_1, \ldots, G_n\}$, where n represents the total number of time periods. Each graph $G_i = (V_i, E_i, A_i)$, where $V_i = \{r_0, r_1, \ldots, r_m\}$ is the node set with r_0 as the root node. Each node represents a microblog in the process of the event propagation, and m is the total number of microblogs involved at time t_i. $E_i = \{e_{st}|s, t = 0, \ldots, m\}$ is the edge set. If there is a forwarding or commenting relationship between the two microblogs r_s and r_t, there is an edge e_{st}. For example, if r_2 is the reposted microblog of r_1, there is a directed edge, $r_1 \rightarrow r_2$, i.e., e_{12}. The adjacency matrix

of G_i can be defined as follows:

$$A_i = \begin{pmatrix} 0 & a_i(1,2) & \cdots & a_i(1,m) \\ a_i(2,1) & 0 & \cdots & a_i(2,m) \\ \vdots & \vdots & \ddots & \vdots \\ a_i(m,1) & a_i(m,2) & \cdots & 0 \end{pmatrix} \tag{2}$$

The element $a_i(s,t)$ of the s-th row and t-th column in A_i is defined as follows:

$$a_i(s,t) = \begin{cases} 1, e_{st} \in E_i \\ 0, else \end{cases} \tag{3}$$

That is, when $e_{st} \in E_i$, the corresponding element value of the s-th row and t-th column of the adjacency matrix A_i is 1, otherwise it is 0.

4.3 Encoding Text Contents

We use text content features as initialization features of nodes in G, so in this section we will describe the method of text content encoding. We adopt Sentence-BERT (SBERT) [27] to encode the text content feature. SBERT is a pre-trained model for encoding text content features. This method is fine-tuned on Bidirectional Encoder Representations from Transformers (BERT) using Siamese and three-level network structure, which improves the traditional sentence embedding method and achieves very good results in multiple application scenarios.

SBERT adds a pooling operation to the output of BERT to derive a fixed sized sentence embedding. The pooling strategy in this paper is MEAN pooling.

The propagation structure graph at time t_i is G_i, the node set is $V_i = \{r_0, r_1, \ldots, r_m\}$, each node r_j is a microblog, and the corresponding text content is w_j, the feature $p_j \in R^{1*d}$ of each node is extracted by the SBERT model:

$$p_j = SBERT(w_j) \tag{4}$$

d is the dimension of the initialization feature vector, and all the node features in the set V_i are combined to obtain the initialization feature matrix T_i of the propagation structure graph G_i:

$$T_i = \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_{m-1} \\ p_m \end{pmatrix} \tag{5}$$

4.4 Spatial Feature Extraction

We employ a two-layer graph convolutional network to extract spatial feature from the dynamic graphs. Graph Convolutional Network (GCN) is an extension of CNN on graph data, which can effectively capture graph features by aggregating the neighborhood information of nodes in the graph.

For each graph in different time period t_i, we use two-layer graph convolutional networks to extract the spatial features. The calculation formula is as follows:

$$H_1^i = \sigma(\hat{A}_i T_i W_0^i) \tag{6}$$

$$H_2^i = \sigma(\hat{A}_i H_1^i W_1^i) \tag{7}$$

where $\hat{A}_i = A_i + I$, I is the identity matrix, $H_1^i \in R^{m*v_0}$, $H_2^i \in R^{m*v_1}$ represent the outputs of the first and second graph convolutional layers (GCL), namely the hidden state. m is the total number of nodes at time t_i, v_0 is the output vector dimension of the first layer, and v_1 is the output vector dimension of the second layer. $W_0^i \in R^{d*v_0}$ and $W_1^i \in R^{v_1*v_1}$ are parameter matrices GCL. $\sigma(\cdot)$ is the activation function, we use the ReLU function here.

Finally, through average pooling, as shown in Eq. (13), we get the final spatial features.

$$S^i = meanpooling(H_2^i) \tag{8}$$

4.5 Temporal Feature Extraction

In this part, we use a recurrent neural network to extract the temporal features of rumor propagation. We take the spatial features $\{S^0, S^1, \ldots, S^n\}$ as input and pass the spatial feature representation S^t to an RNN unit.

This article compares three different RNN units: basic RNN, LSTM, and GRU through experiments, and finally finds that GRU has the best effect. Its calculation formula is as follows:

$$r_t = \sigma(W_{ir}S^t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \tag{9}$$

$$z_t = \sigma(W_{iz}S^t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \tag{10}$$

$$n_t = tanh(W_{in}S^t + b_{in} + r_t \odot (W_{hn}h_{(t-1)} + b_{hn})) \tag{11}$$

$$h_t = (1 - z_t) \odot n_t + z_t \odot h_{(t-1)} \tag{12}$$

where h_t is the hidden state at time t, $h_{(t-1)}$ is the hidden state at time $t-1$. r_t, z_t, n_t are the reset, update, and new gates, respectively. σ is the sigmoid function, and $\odot$ is the Hadamard product.

The output h_n of the last unit is the result of fusing spatial and temporal features. Finally, we concatenate the source Weibo text feature representation with h_n to enhance the feature representation of rumors:

$$D = concat(SBERT(w_0), h_n) \tag{13}$$

4.6 Classification and Training

The predicted label $\hat{y}$ of the event is calculated by multiple fully connected layers and a $softmax$ layer:

$$\hat{y} = softmax(FC(D)) \tag{14}$$

where $\hat{y} \in R^{(1*K)}$ is a probability vector, K is the number of classes, and the value $\hat{y}_i$ of each element of $\hat{y}$ represents the probability that the event belongs to the corresponding class.

We train all parameters in the model by minimizing the cross-entropy between the predicted result and the ground truth Y of all events, and add the L_2 regular term during the training process to avoid the overfitting problem. The loss function L is defined as:

$$L = \sum_{|C|} \sum_{i \in \{0,1\}} -y_i log\widehat{y_i} + \beta L_2 \tag{15}$$

where β is the coefficient of the regular term and $|C|$ is the total number of events.

4.7 Explanation

Although our rumor detection module can accurately identify rumors, effective curbing of their spread is difficult without providing reasonable explanations. To address this, we utilize a LLM to offer coherent explanations for the rumors. However, in real social scenarios, the number of comments can be quite large, while LLMs have limitations on input length. Furthermore, LLMs often struggle to handle excessively long or redundant information [28,29].

In recent research on LLMs [30–32], complex tasks are typically decomposed into a series of simpler tasks, which significantly enhances the ability of LLMs to handle complexity. Accordingly, we delineate the prompts into a **Chain-of-Propagation** (CoP), facilitating easier reasoning for the LLMs. Given that comments within a certain timeframe often share similar sentiments, we construct the CoP based on the chronological order of comment timestamps, as illustrated in Fig. 4.

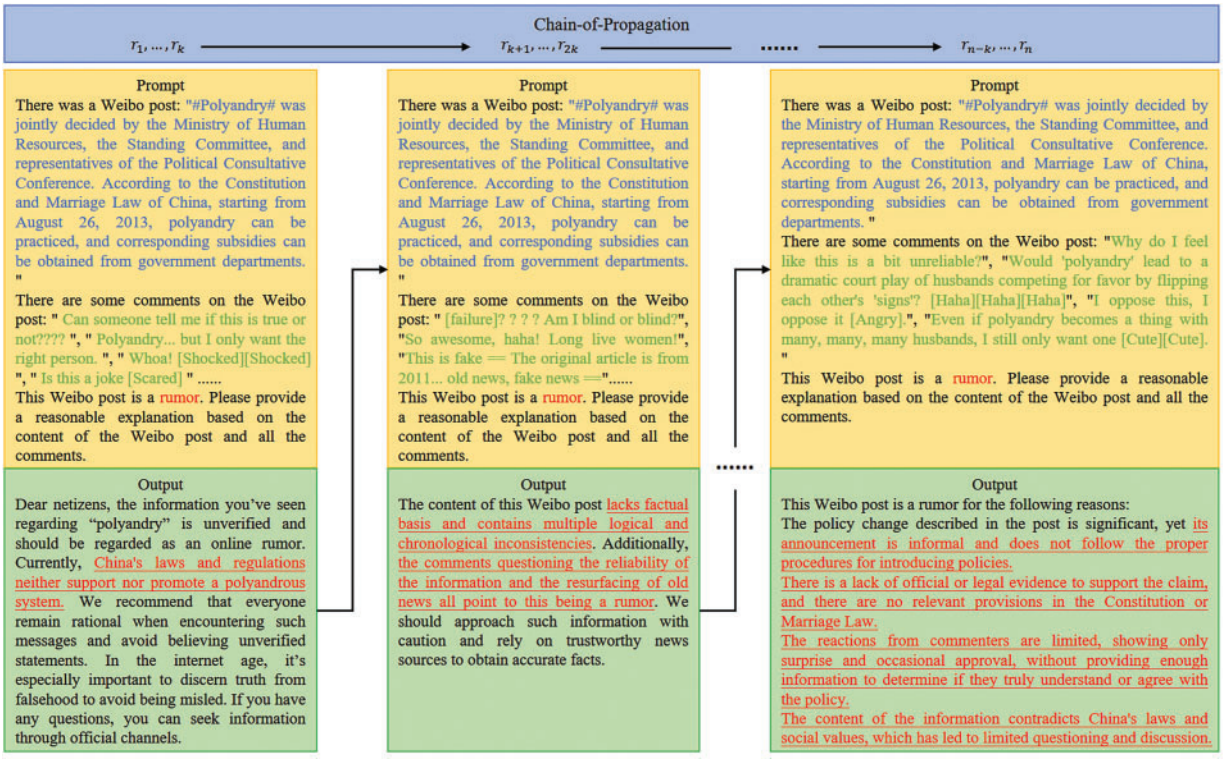

Figure 4: Schematic diagram of CoP

We divide all the comments $r_1, r_2, \ldots, r_n$ in a given event c_i into m groups, each containing k comments. In the inference prompt for the LLMs, we select only one group at a time as the input (as shown by the green text in the prompt of Fig. 4). After m rounds of inference, we obtain the final result. The inference steps within the CoP operate in the same session of the LLM, allowing subsequent steps to reference the results of prior steps. We utilize the output from the final reasoning step as the conclusive result, as it aggregates information from all preceding steps. In each round of the inference prompt, the source tweet r_0 of the event is also included (as shown by the blue text in the prompt of Fig. 4), along with the rumor detection result (as shown by the red text in the prompt of Fig. 4).

5 Experiment

5.1 Datasets

We benchmark our proposed method against state-of-the-art baselines using two public datasets: WEIBO [33] and PHEME [34]. The WEIBO dataset, compiled from Xinhua News Agency and Weibo, comprises a rich collection of data. Similarly, the PHEME dataset is an aggregation of content centered around five distinct breaking news stories, with each story featuring a collection of associated posts. Adhering to the methodology outlined in [35], both the WEIBO and PHEME datasets are segmented into training and testing subsets with an 8:2 ratio. Table 1 shows the statistics of the datasets.

Table 1: Statistics of datasets

Statistic	Weibo	PHEME
# of events	4657	5748
# of rumors	2345	2094
# of non-rumors	2312	3654
Avg. # of posts/event	804	16
Max # of posts/event	59,318	346
Min # of posts/event	10	1

5.2 Experiment Setup

Our hardware environment is configured with a Hygon C86 3285 8-core processor, 128 GB of memory, and an NVIDIA RTX A6000 graphics card. We implemented our rumor detection method using the PyTorch framework, with parameter optimization using the Adam algorithm. In our model, we utilized the pre-trained SBERT: all-MiniLM-L6-v2 and selected the Llama-3-8B [36] large language model as the rumor result interpreter.

In terms of model parameter settings, the output feature dimension of SBERT is 512. The GCL layer uses two layers of GCN, with the first layer having an input feature dimension of 512 and an output feature dimension of 256, while the second layer has an output feature dimension of 50. In the RNN layer, we select a single GRU layer as the temporal feature extraction model, with a hidden layer feature dimension of 100. The classifier has an input feature dimension of 100 and an output dimension corresponding to the number of classes, which is 2.

We frequently employ Accuracy as the primary evaluation metric for binary classification tasks, including the detection of fake news. Nonetheless, the reliability of Accuracy is significantly undermined when dealing with datasets that exhibit class imbalance. To address this limitation, our experimental framework incorporates a suite of complementary metrics alongside Accuracy. Specifically, we introduce Precision, Recall, and the F1 score to provide a more nuanced and comprehensive assessment of our model's performance in the context of rumor detection.

5.3 Baselines

We compare the following baseline models with our model:

- SVM-TS [37]: SVM-TS detects fake news using heuristic rules and a linear SVM classifier.
- CNN [9]: This CNN-based method learns feature representations for early misinformation detection by analyzing posts in fixed-length sequences.

- GRU [12]: The GRU model, an RNN variant, excels at capturing contextual information from related posts over time.
- TextGCN [38]: TextGCN uses graph convolutional networks to improve word and document embedding, viewing the corpus as a heterogeneous graph.
- EANN [39]: EANN, a GAN-based model, extracts event-invariant features for new event detection.
- RumorGAN [11]: RumorGAN generates conflicting or uncertain signals to strengthen the discriminator (GRU), enabling it to learn more robust representations of rumors.
- GACL [16]: A GNN-based model using adversarial and contrastive learning to encode global propagation, resist noise and adversarial samples, and capture event-invariant features.
- BiGCN [14]: BiGCN is a model based on GCN, which can embed propagation structure and diffusion structure at the same time.

5.4 Comparative Experimental Results and Analysis

Table 2 presents the experimental outcomes for our approach and the baseline methods. Key observations include:

Table 2: The performance results of the comparison methods on Weibo and PHEME

Method	Weibo					PHEME			
	Class	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
SVM-TS	R	0.64	0.741	0.573	0.646	0.639	0.546	0.576	0.560
	N		0.651	0.798	0.711		0.729	0.705	0.717
CNN	R	0.74	0.736	0.756	0.744	0.779	0.732	0.606	0.663
	N		0.747	0.723	0.735		0.799	0.875	0.835
GRU	R	0.702	0.671	0.794	0.727	0.832	0.782	0.712	0.745
	N		0.747	0.609	0.671		0.855	0.896	0.865
TextGCN	R	0.787	0.925	0.573	0.727	0.828	0.775	0.735	0.737
	N		0.712	**0.985**	0.827		0.827	0.828	0.828
EANN	R	0.782	0.827	0.697	0.756	0.681	0.685	0.664	0.694
	N		0.752	0.863	0.804		0.701	0.750	0.747
RumorGAN	R	0.867	0.906	0.815	0.858	0.783	0.725	0.772	0.748
	N		0.826	0.917	0.869		0.845	0.794	0.818
GACL	R	0.915	0.928	0.947	0.937	0.850	0.801	0.750	0.774
	N		0.872	0.912	0.892		0.871	0.901	0.885
BiGCN	R	0.919	0.936	0.952	0.944	0.847	0.820	0.787	**0.803**
	N		0.901	0.894	0.897		0.862	0.883	0.872
Ours	R	**0.941**	**0.937**	**0.957**	**0.947**	**0.868**	**0.832**	**0.769**	0.799
	N		**0.946**	0.922	**0.931**		**0.884**	**0.919**	**0.901**

(1) Across all datasets, SVM-TS underperforms, suggesting that manually engineered features may be inadequate for fake news detection.

(2) Deep learning models (CNN, GRU, RumorGAN, GACL) surpass SVM-TS, highlighting their advantages over conventional techniques.

(3) Our method is superior to other methods in both data sets, which proves the effectiveness of feature fusion in temporal domain and spatial domain.

5.5 Ablation Study

We did an ablation study to see whether each module contributes to the model and which modules contribute more. Our model's main components include SBERT, GCN, and GRU. Based on the complete model, we systematically remove the module and compare their changes in accuracy, precision, recall, F1 value. Fig. 5 shows the experiment results.

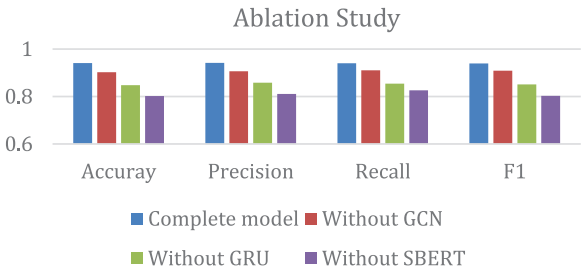

Figure 5: Ablation study results

The ablation study reveals that the performance of the model is significantly reduced when any of the components (SBERT, GCN, or GRU) are removed. This confirms that each module makes a meaningful contribution to the model's overall ability to detect rumors accurately.

- Full Model. The complete model, which integrates SBERT, GCN, and GRU, achieved the highest performance across all metrics. This demonstrates that combining textual, spatial, and temporal features results in a robust and accurate rumor detection system.
- Without SBERT. Removing the SBERT module resulted in the greatest performance drop, underscoring the importance of textual information in detecting rumors. Without SBERT, the model struggled to differentiate between rumor and non-rumor text effectively.
- Without GRU. When the GRU module, responsible for extracting temporal features, was removed, the model's ability to capture the dynamic evolution of rumors over time was significantly hindered. This resulted in reduced recall and F1 scores, demonstrating that the temporal dimension is crucial for effective rumor detection, especially in understanding how rumors evolve over time.
- Without GCN. The removal of the GCN module also led to a noticeable decline in performance, particularly in accuracy and recall. This suggests that the GCN's ability to capture spatial features in the propagation structure of rumors is essential for improving the model's ability to identify and track the spread of rumors in the network.

5.6 Impact of Parameters

Next, we evaluated the impact of several key parameters in our method, with validation results shown in Fig. 6. We selected three parameters for validation: RNN type, learning rate, and number of iterations.

The choice of these hyperparameters was guided by several considerations. First, the RNN type (such as LSTM, GRU, or Simple-RNN) was chosen based on its ability to capture the temporal dynamics of

rumor propagation, with LSTM and GRU generally offering better performance in handling long-term dependencies in time series data. Second, the learning rate was selected as a crucial factor influencing the convergence speed and stability of the model. Lastly, the number of iterations was determined by balancing model training time and accuracy. Too few iterations may result in underfitting, while too many could lead to overfitting or unnecessary computational overhead.

For these three parameters, we used cross-validation to assess performance on the Weibo dataset, which served as a case study in our experiments. To ensure the robustness of the results, we kept the other parameters fixed at their optimal values while adjusting one parameter at a time. This allowed us to isolate the effects of each parameter and find the best configuration for the task at hand. The final hyperparameter values were selected based on the validation performance, ensuring the model achieved the best possible balance between accuracy and computational efficiency.

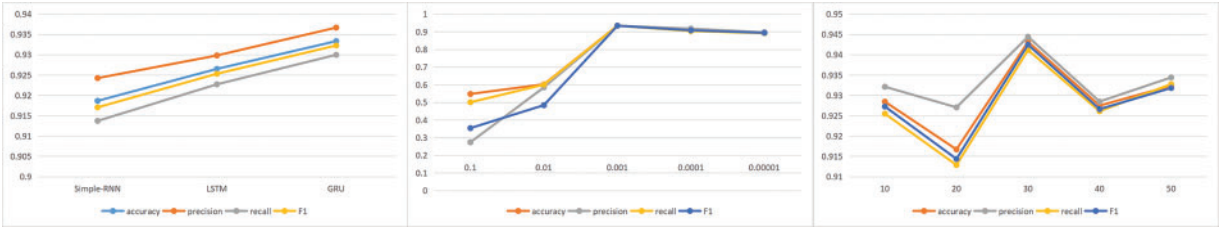

Figure 6: Impact of RNN types, learning rate and epoch

In Fig. 6, the x-axis represents different types of RNNs, learning rates, and iteration epochs, while the y-axis displays the values of accuracy, precision, recall, and F1 score. The curves in different colors represent the variation trends of different evaluation metrics.

5.7 Case Study of Explanation

After obtaining the model's classification results, we used a LLM to explain these results, particularly for events classified as rumors by the model. We chose Llama-3-8B as the interpreter and selected a real tweet from the Weibo dataset as an example. The tweet's main content discussed China's supposed plan to implement a "polyandry" system. Table 3 shows the prompts and the outputs.

Table 3: The performance results of the comparison methods on PHEME

Prompt of Step 1	There was a Weibo post: "#Polyandry# was jointly decided by the Ministry of Human Resources, the Standing Committee, and representatives of the Political Consultative Conference. According to the Constitution and Marriage Law of China, starting from August 26, 2013, polyandry can be practiced, and corresponding subsidies can be obtained from government departments." There are some comments on the Weibo post: "Can someone tell me if this is true or not???? ", " Polyandry... but I only want the right person. ", " Whoa! [Shocked][Shocked] ", " Is this a joke [Scared] "..... This Weibo post is a rumor. Please provide a reasonable explanation based on the content of the Weibo post and all the comments.

(Continued)

Table 3 (continued)

Output of Step 1	Dear netizens, the information you've seen regarding "polyandry" is unverified and should be regarded as an online rumor. Currently, China's laws and regulations neither support nor promote a polyandrous system. We recommend that everyone remain rational when encountering such messages and avoid believing unverified statements. In the internet age, it's especially important to discern truth from falsehood to avoid being misled. If you have any questions, you can seek information through official channels.
Prompt of Step 2	There was a Weibo post: "#Polyandry# was jointly decided by the Ministry of Human Resources, the Standing Committee, and representatives of the Political Consultative Conference. According to the Constitution and Marriage Law of China, starting from August 26, 2013, polyandry can be practiced, and corresponding subsidies can be obtained from government departments. " There are some comments on the Weibo post: " [failure]? ? ? ? Am I blind or blind?", "So awesome, haha! Long live women!", "This is fake == The original article is from 2011... old news, fake news =="..... This Weibo post is a rumor. Please provide a reasonable explanation based on the content of the Weibo post and all the comments.
Output of Step 2	The content of this Weibo post lacks factual basis and contains multiple logical and chronological inconsistencies. Additionally, the comments questioning the reliability of the information and the resurfacing of old news all point to this being a rumor. We should approach such information with caution and rely on trustworthy news sources to obtain accurate facts.
Final output	This Weibo post is a rumor for the following reasons: The policy change described in the post is significant, yet its announcement is informal and does not follow the proper procedures for introducing policies. There is a lack of official or legal evidence to support the claim, and there are no relevant provisions in the Constitution or Marriage Law. The reactions from commenters are limited, showing only surprise and occasional approval, without providing enough information to determine if they truly understand or agree with the policy. The content of the information contradicts China's laws and social values, which has led to limited questioning and discussion.

We employed CoP method, where multiple steps of prompts were used to obtain the final output. The selected example contained 143 reposts or comments, with 113 remaining after removing empty entries. We set the number of comments included in each reasoning step to K = 5, resulting in a total of 23 reasoning steps. Table 3 includes the first two reasoning steps' prompts and outputs, as well as the final model output.

In Table 3, the sentences highlighted in blue represent the tweet, the green sentences represent comments, and the red sentences provide evidence that the event is a rumor. From the red text, we can observe that as the reasoning steps progressed, the amount of evidence in the output increased, while irrelevant sentences decreased.

In the first reasoning step, only one sentence was able to explain why the tweet was a rumor, and it didn't consider the comment information. However, in the second reasoning step, there were two sentences serving as evidence for the tweet being a rumor, and the model began to take comments into account, using them to explain why the tweet was classified as a rumor. After 23 reasoning steps, the final output provided a comprehensive explanation of the rumor, covering four key points: 1) It did not follow proper policy-making procedures; 2) It lacked legal basis; 3) There was limited reaction in the comments; 4) It contradicted China's social values.

The advantage of this method lies in its ability to automatically generate highly persuasive explanatory text for rumors, without requiring manual refinement of language. Additionally, it is simple to use, requiring only carefully designed prompts. However, its limitation is that it cannot reveal the internal mechanisms of the model's prediction algorithm.

6 Conclusions

This paper proposes a novel rumor detection method that integrates temporal and spatial features, and builds upon this by introducing Llama-3-8B to explain the detection results. Different from the existing methods, our method can detect rumors more accurately and automatically generate text to explain the detection results. Through experiments on the Weibo and PHEME datasets, we demonstrate that this method outperforms existing models. Ablation studies confirm the importance of each model component, particularly the SBERT module, highlighting the critical role of textual information in rumor detection. Case studies show that our approach effectively explains the detection results. Although our model performs well, we recognize the importance of continuous improvement to address the issue of social media rumors. Future research will focus on expanding the model's adaptability and exploring its application across various data sources and emerging platforms. Our work represents a significant step forward in the fight against online rumors, laying a solid foundation for the development of more complex detection systems.

Acknowledgement: Thanks to the project financial support provided by Zhejiang Provincial Department of Education and the experimental environment provided by Zhejiang Police College.

Funding Statement: This work was supported by General Scientific Research Project of Zhejiang Provincial Department of Education (Y202353247).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Nanjiang Zhong; data collection: Nanjiang Zhong; analysis and interpretation of results: Nanjiang Zhong, Xinchen Jiang, Yuan Yao; draft manuscript preparation: Nanjiang Zhong. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Nanjiang Zhong, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Allcott H, Gentzkow M. Social media and fake news in the 2016 election. J Econ Perspect. 2017;31(2):211–36. doi:10.1257/jep.31.2.211.
2. Islam MS, Sarkar T, Khan SH, Kamal A-HM, Hasan SM, Kabir A, et al. COVID-19-related infodemic and its impact on public health: a global social media analysis. Am J Trop Med Hyg. 2020;103(4):1621. doi:10.4269/ajtmh.20-0812.

3. Liu X, Nourbakhsh A, Li Q, Fang R, Shah S. Realtime rumor debunking on twitter. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management; 2015; New York, NY, USA: Association for Computing Machinery. p. 1867–70.
4. Zhao Z, Resnick P, Mei Q. Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide; 2015; Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. p. 1395–405.
5. Castillo C, Mendoza M, Poblete B. Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web; 2011; New York, NY, USA: Association for Computing Machinery. p. 675–84.
6. Cheng M, Nazarian S, Bogdan P. VRoC: variational autoencoder-aided multi-task rumor classifier based on text. In: Proceedings of the Web Conference 2020; 2020; New York, NY, USA: Association for Computing Machinery.
7. Ruchansky N, Seo S, Liu Y. CSI: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 2017; New York, NY, USA: Association for Computing Machinery. p. 797–806.
8. Yu F, Liu Q, Wu S, Wang L, Tan T. A convolutional approach for misinformation identification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017 August 19–25; Melbourne, Australia. p. 3901–3907.
9. Ma J, Gao W, Wong KF. Detect rumor and stance jointly by neural multi-task learning. In: Companion Proceedings of the the Web Conference 2018; 2018; Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. p. 585–93.
10. Pattanaik B, Mandal S, Tripathy RM. A survey on rumor detection and prevention in social media using deep learning. Knowl Inf Syst. 2023;65. 10:3839–80.
11. Ma J, Gao W, Wong KF. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In: The World Wide Web Conference; 2019; New York, NY, USA: Association for Computing Machinery.
12. Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong K-F, et al. Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16; 2016; New York, NY, USA: AAAI Press. p. 3818–24.
13. Li R, Jiang Z, Gao S, Yang W. Incorporating neural point process-based temporal feature for rumor detection. In: International Conference on Combinatorial Optimization and Applications; 2023; Cham, Switzerland: Springer Nature. p. 419–30.
14. Bian T, Xiao X, Xu T, Zhao P. Rumor detection on social media with bi-directional graph convolutional networks. Proc AAAI Conf Artif Intell. 2020;34:549–56.
15. Naumzik C, Feuerriegel S. Detecting false rumors from retweet dynamics on social media. In: Proceedings of the ACM Web Conference 2022; 2022; New York, NY, USA: Association for Computing Machinery. p. 2798–809.
16. Sun T, Qian Z, Dong S, Li P, Zhu Q. Rumor detection on social media with graph adversarial contrastive learning. In: Proceedings of the ACM Web Conference 2022; 2022; New York, NY, USA: Association for Computing Machinery. p. 2789–97.
17. Ma J, Gao W, Wong KF. Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017); 2017 Jul 30–Aug 4; Vancouver, BC, Canada: Association for Computational Linguistics; p. 708–17.
18. Kwon S, Cha M, Jung K. Rumor detection over varying time windows. PLoS One. 2017;12(1):e0168344. doi:10.1371/journal.pone.0168344.
19. Wu K, Yang S, Zhu KQ. False rumors detection on Sina Weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering; 2015. p. 651–62.
20. Kumar S, Hamilton WL, Leskovec J, Jurafsky D. Community interaction and conflict on the web. In: Proceedings of the 2018 World Wide Web Conference; 2018; Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. p. 933–43.
21. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359(6380):1146–51. doi:10.1126/science.aap9559.

22. Chen C, Shu K. Combating misinformation in the age of LLMs: opportunities and challenges. arXiv preprint arXiv:2311.05656. 2023.
23. Ivan V, Matúš P, Ivan S, Robert M, Dominik M, Maria B. Disinformation capabilities of large language models. arXiv preprint arXiv:2311.08838. 2023.
24. Chen C, Shu K. Can LLM-generated misinformation be detected? arXiv preprint arXiv:2309.13788. 2024.
25. Yang C, Zhang P, Qiao W, Gao H, Zhao J. Rumor detection on social media with crowd intelligence and ChatGPT-assisted networks. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023; Singapore. p. 5705–17.
26. Hu B, Sheng Q, Cao J, Shi Y, Li Y, Wang D, et al. Bad actor, good advisor: exploring the role of large language models in fake news detection. Proc AAAI Conf Artif Intell. 2024;38(20):22105–13.
27. Reimers N. Sentence-BERT: sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084. 2019.
28. Huang Y, Xu J, Jiang Z, Lai J, Li Z, Yao Y, et al. Advancing transformer architecture in long-context large language models: a comprehensive survey. arXiv preprint arXiv:2311.12351. 2023.
29. Xie W. Analysis of the reasoning with redundant information provided ability of large language models. arXiv preprint arXiv:2310.04039. 2023.
30. Besta M, Blach N, Kubicek A, Gerstenberger R, Gianinazzi L, Gajda J, et al. Graph of thoughts: solving elaborate problems with large language models. AAAI Tech Track Nat Lang Process I. 2024;38(16):17682–90.
31. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. In: Advances in Neural Information Processing Systems (NeurIPS 2023); 2023 Dec 10–16; New Orleans, LA, USA.
32. Liu Q, Tao X, Wu J, Wu S, Wang L. Can large language models detect rumors on social media? arXiv preprint arXiv:2402.03916. 2024.
33. Jin Z, Cao J, Guo H, Zhang Y, Luo J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia; 2017; New York, NY, USA: Association for Computing Machinery. p. 795–816.
34. Zubiaga A, Liakata M, Procter R. Exploiting context for rumour detection in social media. In: International Conference on Social Informatics; 2017; Berlin/Heidelberg, Germany: Springer. p. 109–23.
35. Qian S, Wang J, Hu J, Fang Q, Xu C. Hierarchical multi-modal contextual attention network for fake news detection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021 Jul; New York, NY, USA: Association for Computing Machinery. p. 153–62.
36. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783. 2024.
37. Ma J, Gao W, Wei Z, Lu Y, Wong K-F. 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; 2015; New York, NY, USA: Association for Computing Machinery. p. 1751–4.
38. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. Proc AAAI Con Arti Intell. 2019;33(1):7370–7.
39. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, et al. EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018; New York, NY, USA: Association for Computing Machinery. p. 849–57.