Article

Trust-Aware Evidence Reasoning and Spatiotemporal Feature Aggregation for Explainable Fake News Detection

Jing Chen *, Gang Zhou *, Jicang Lu, Shiyu Wang, Shunhang Li

Posted Date: 17 April 2023

doi: 10.20944/preprints202304.0387.v1

Keywords: fake news detection; explainable machine learning; spatiotemporal structure; social network

Article

Trust-Aware Evidence Reasoning and Spatiotemporal Feature Aggregation for Explainable Fake News Detection

Jing Chen *, Gang Zhou *, Jicang Lu, Shiyu Wang and Shunhang Li

State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, 450001, China; lujicang@sina.com; share_wind@163.com; baikal_sh@163.com

* Correspondence: cathysilense@126.com(J.C.); gzhou@126.com (G.Z.)

Abstract: Fake news detection has become a significant topic based on the fast-spreading and detrimental effects of such news. Many methods based on deep neural networks learn clues from claim content and message propagation structure or temporal information, which have been widely recognized. However, such models (i) ignore the fact that information quality is uneven in propagation, which makes semantic representations unreliable. (ii) Most models do not fully leverage spatial and temporal structure in combination. (iii) Finally, internal decision-making processes and results are non-transparent and unexplained. In this study, we develop a trust-aware evidence reasoning and spatiotemporal feature aggregation model for more interpretable and accurate fake news detection. Specifically, we first design a trust-aware evidence reasoning module to calculate the credibility of posts based on a random walk model to discover high-quality evidence. Next, from the perspective of spatiotemporal structure, we design an evidence-representation module to capture the semantic interactions granularly and enhance the reliable representation of evidence. Finally, a two-layer capsule network is designed to aggregate the implicit bias in evidence while capturing the false portions of source information in a transparent and interpretable manner. Extensive experiments on two benchmark datasets indicate that the proposed model can provide explanations for fake news detection results, as well as can achieve better performance, boosting 3.5% in F1-score on average.

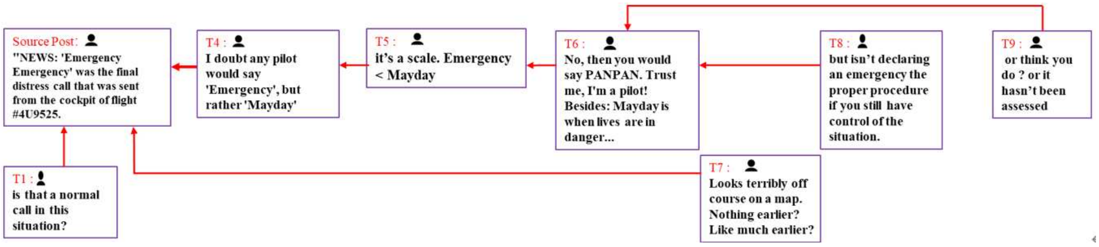
Keywords: fake news detection; explainable machine learning; spatiotemporal structure; social network

1. Introduction

Social media has become a main platform for users to exchange and share messages based on its openness and anonymity. However, based on its low barrier of entry, and rapid provision and dissemination of online news, it also provides a hotbed for the rapid dissemination of disinformation such as fake news. Fake news threatens the security of cyberspace, as well as affects public opinion regarding major social events [1]. This can seriously interfere with personal cognition, cause people to make incorrect decisions, and even exert a serious negative influence on the political order and economy in the real world. For instance, during the 2016 US presidential election, various types of fake news were more popular and widespread on Facebook than sources of accurate news, which affected voter cognition, even changing the proportion of people supporting different parties and having a significant impact on the fairness of the election [2]. In the "information plague" accompanying the COVID-19 pandemic in 2020, many news reports with misleading content spread through social media, leading to socioeconomic disorder and reduced effectiveness of national epidemic prevention measures [3]. Therefore, constructing a detection model to curb fake news on social media has important theoretical value and practical significance for maintaining national security and social stability.

Existing fake news detection methods has two main classes, namely content-based and social context-based methods [4]. Specifically, content-based methods are dedicated to modeling the text content of news. For instance, early approaches considers linguistic features [5–7], topics [8] and emotional features [9,10] in a manual way, and more recent methods extract the higher-level and implicit semantic information from news' text contents by neural networks [11–14]. These methods have achieved huge success on regular fake news detection. However, compared to news articles, the information published on social media is short and non-standard in format, and contains less effective information with more noise leading to the issue of sparse semantics [15]. Therefore, it is difficult for previous models which process long and standard text contents of regular news to extract key semantic features for detection from the short news posted on social media due to the semantic sparsity. To alleviate the semantic sparsity issue, recent studies have tried to introduce additional information source (e.g., social context information). Social context information (e.g., interactions between users and news story), providing abundant reference information, boasts great potential in alleviating the task, leading to the social context-based methods. Concretely, they can be further divided into posts-based and propagation-based methods. Posts-based methods utilize user opinions regarding relevant content to help fake news detection, by modeling the semantic interactions between source information and user comments [16–18]. Motivated to the posts-based methods, propagation-based methods further consider the local characteristics of semantic interactions in message propagation, to capture and distinguish user views regarding source information and comments in a fine-grained manner. Specifically, they model the information propagation process as a graph structure and use graph neural networks to aggregate spatial neighborhood information and learn the high-level spatial structure semantic representations [19–23].

Although the methods discussed above have improved the detection performance, they still have some limitations. First, based on the openness and low barrier to entry of social media, there may be artificial accounts [24] attempting to publish incorrect information to affect public opinion during message propagation. When aggregating neighborhood information, existing models treat all information equally, which may introduce noise and render semantic representations unreliable [25]. Therefore, it is necessary to mitigate the impact of noise on model detection by calculating the credibility of comments. Second, existing propagation-based research methods mainly model spatial propagating structure characteristics without considering the dynamic evolution of posts over time. As shown in Figure 1, both comments T_8 and T_9 are replying to comment T_6 with the same spatial structure characteristics, but from the perspective of time, there are clear differences between them (T_8 is released earlier than T_9, meaning users may already be affected by T_8 when T_9 is released). Recent studies [26–28] have demonstrated that temporal structure features can capture the dynamic evolution of information in a more fine-grained manner and promote the early detection performance. Spatial and temporal structures depict the evolution of news messages from the perspectives of information interaction network and temporal message propagation respectively, which are complementary. Therefore, it is necessary to consider both the temporal neighborhood structure characteristics and spatial neighborhood structure characteristics of information. Additionally, existing methods focus on using deep learning models to integrate more external information and automatically mine hidden features to improve fake news detection performance. While, with the model complexity increasing, the decision-making process within a model has become more difficult to explain and verify. A psychological research [29] has shown that the spreading power of false information is closely related to the importance and fuzziness of events. Therefore, it is typically insufficient to simply mark information as false. A model must also automatically provide a judgment basis to enhance interpretability.



"NEWS: 'Emergency Emergency' was the final distress call that was sent from the cockpit of flight #4U9525.

I doubt any pilot would say 'Emergency', but rather 'Mayday'

it's a scale. Emergency < Mayday

No, then you would say PANPAN. Trust me, I'm a pilot! Besides: Mayday is when lives are in danger...

but isn't declaring an emergency the proper procedure if you still have control of the situation.

or think you do ? or it hasn't been assessed

is that a normal call in this situation?

Looks terribly off course on a map. Nothing earlier? Like much earlier?

Figure 1. Schematic diagram of the spatiotemporal structure of information dissemination.

To alleviate the problems discussed above, we designed a **T**rust-aware evidence **R**easoning and **S**patiotemporal feature **A**ggregation (TRSA) model to discover evidence for interpretable fake news detection. Specifically, we first design a trust-aware evidence reasoning module to calculate the credibility of posts based on a random walk model to discover high-quality posts as evidence. Then, considering the credibility of evidence, we design an evidence representation module based on spatiotemporal structure to aggregate the spatiotemporal neighborhood characteristics of message propagation and enhance the reliable representation of evidence. Finally, we detect fake news by aggregating the implicit bias of evidence in source information based on a capsule network. Specifically, we first model the semantic interactions between evidence and source information to capture the controversial points (false portions) of source information and form an evidence capsule. We then aggregated the implicit bias of each evidence capsule from the source information through a dynamic routing mechanism. This study makes the following contribution and innovations.

- We develop a transparent and highly interpretable neural structure reasoning model that incorporates a random walk model and capsule network structure into the processes of evidence reasoning and aggregation, respectively, which not only provides reliable evidence for fake news detection, but also enhances the transparency of the model reasoning process.
- Our evidence representation module can capture the semantic interactions between posts in a fine-grained manner based on the spatiotemporal structure of message propagation to enrich the semantic representation of posts (source information or comments).
- The designed evidence aggregation module automatically captures the false portions of source information while aggregating the implicit bias of the evidence in source information.
- Extensive experiments on public datasets illustrate that TRSA achieves more promising performance than previous state-of-the-art approaches, as well as provide interpretations for fake news detection results.

2. Related Work

In essence, fake news detection can be considered as a classification problem. Specifically, for a given input, a model should output a true or fake label. Therefore, many researchers have focused on mining various characteristic signals of fake news such as text, publishing users, participating users, and communication networks, and have attempted to develop better methods for feature extraction.

Content-based approaches. Early methods based on content mainly focused on manually extracting various lexical, grammatical, or topical features, and using traditional machine learning methods to detect fake news. For example, kwon et al. [30] found that emotion features are valuable for constructing false information classification models (including positive emotion words, negative words, and cognitive behavior words). On this basis, a time series model was designed to capture the key language differences between true and false information. Potthast et al. [31] used different writing styles to identify false statements. Ito et al.[8] introduced a potential Dirichlet distribution topic model for Twitter reliability evaluation. To avoid handcrafted feature engineering and automatically capture the deep hidden semantic features of text, various deep neural network models have been developed. Ma et al. [32] used recurrent neural networks to mine high-level hidden features in information for fake news detection. Wang et al. [33] applied a pre-training model to detect false information and achieved good results. Hu et al. [34] constructed a heterogeneous graph containing topic, sentence, and entity information to represent a news document and developed a novel network to distinguish fake news. These methods are significantly effective for distinguishing false news articles. However, news published on social media is short leading to the issue of sparse semantics. Therefore, the detection performance of these models are significantly reduced.

Social context-based approaches. Social media is essentially a heterogeneous graph that includes users, posts, and other entities, as well as forwarding, commenting, and other relationships. Therefore, we can integrate social context information from different perspectives to perform fake news detection tasks. Social context-based approaches can be further grouped into posts- and propagation-based methods. **Posts-based methods** mainly rely on user reviews on social media, which can provide useful indicators for distinguishing false information. Therefore, user social responses in terms of emotions, opinions, or stances can be extracted through comments to optimize model detection performance. Wu et al. [17,18] used multitask learning and co-attention network to capture both source information and comments jointly to improve task performance. Zhang et al. [35] hypothesized that fake news can often attract attention and arouse or activate emotions. Therefore, news comments (i.e., social emotions) from a crowd should not be ignored. The shortcoming of such models is that they overlook the obvious local characteristics of social media information interactions. **Propagation-based methods** mainly construct isomorphic or heterogeneous information propagation network modeling interactions between posts or users and mining the information propagating structural characteristics for evaluating the authenticity of one claim. Yuan et al. [19]proposed a novel attention network that jointly encodes local semantic information and global propagation structure information for false information detection. Bian et al. [20] devised a two-layer graph neural network (BiGCN) model to capture the bidirectional propagating structure of information. Although these models optimize performance by mining information dissemination structure features, they rely too heavily on the feature extraction performance of graph neural networks. As the complexity of a model increases, the decision-making process within the model becomes more difficult to explain and verify.

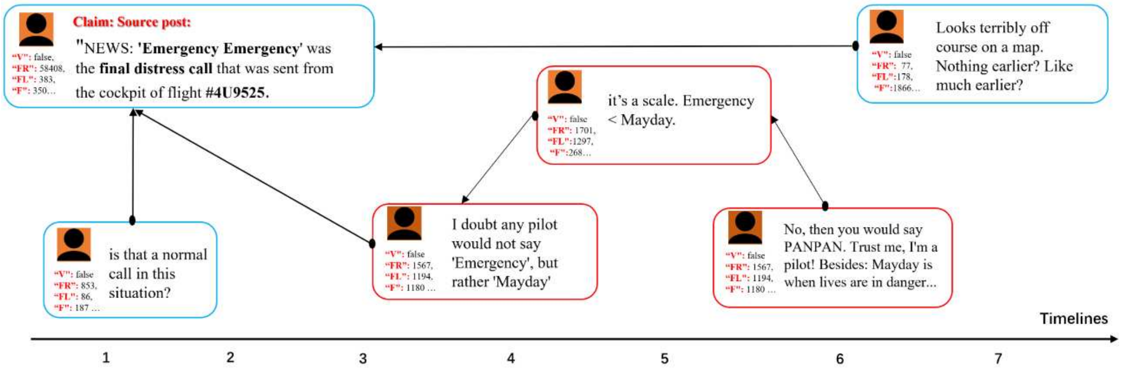
Interpretable machine learning. Our study is also correlated with interpretable machine learning, which mainly focuses on two aspects: the explanation of models and explanation of results. The explanation of models primarily relies on probability graph models or knowledge graphs technologies. For example, Shi et al. [36] proposed a KG-based method to verify facts through predicate paths. Ciampaglia et al. [37] hypothesized that the shortest path between concept nodes could be determined by defining appropriate semantic proximity indicators on a knowledge graph, which can effectively approximate human fact checking. Most automatic fact-checking methods require a knowledge base and must be updated regularly to ensure that the knowledge base remains current. However, it is difficult to handle new topics such as symptoms of COVID-19 at an early stage. The explanation of results primarily is dedicated to visualizing the attention distribution in model decision process. For example, Chen et al. [38] found that words reflecting negative emotions have higher attention weights than words related to events by visualizing the attention weights of their model. Wu et al. [18] depicted semantic interaction between high-quality comments and claim with a co-attention mechanism and found that the attention weights of evidence-related words are higher than that of other words.

In contrast to previous studies, we first calculate the credibility of nodes in an information dispersion network from the perspectives of user authority and communication relationships, and filter highly informative evidence from comments. Second, by incorporating node credibility, we capture the spatiotemporal structure characteristics of nodes from the perspectives of time and space of information propagation and enhance the reliable representation of evidence. Finally, a capsule network is used to model the implicit bias of evidence relative to source information in a transparent manner to enhance the interpretability of the model.

3. Problem Statement

Let Ψ={S_1,S_2,...,S_n} be the source information to be detected and $U = \{u_1, u_2, \ldots, u_l\}$ be a user collection on social media. Each $S_i \in \Psi$ consists of a sequence of l_i tokens $\{w_1, w_2, \ldots, w_{li}\}$, where each token $w_{li} \in R^d$ is a d-dimensional vector denoting the token feature. Each $u_i \in R$ represents the authority of user i calculated based on multiple meta-data features, including whether account verified, geolocation information and homepage introduction exist, and numbers of fans, friends, and favorites. The specific calculation process is provided in Appendix A.

When a news story S_i is posted, it causes users to discuss the story and generate comments or forwarded information. We describe information propagation process from the perspective of temporal and spatial structures, as shown Figure 2. The spatial structure reflects user-content interaction (e.g., commenting, forwarding) in message propagation, while the temporal structure associate with a series of posts (e.g., comments or forwarded messages) over time. We denote the temporal structure as $P(S_i) = \{(c_0, t_0), \ldots, (c_j, t_j), \ldots\}$, where c_j is a d-dimensional vector representing the post (comment or forwarded) content at time j in the propagation of information S_i and t_j is the time at which post c_j is generated. c_0 denotes S_i semantic feature. The spatial structure is denoted as $G(S_i)=<V, E>$, where $G(S_i)$ represents propagation graph of news S_i. V is the node collection of $G(S_i)$ denoting posts in source information propagation. Each node in V is defined as $v_j = (u_j, c_j) \in V$, where $u_j \in R$ represents the authority of the user who posted post j, $c_j \in R^d$ characterizes the post content. E denotes the edge collection describing the association relationship between nodes in $G(S_i)$. If c_j is a comment or forwarded message to c_i, a directed edge from node i to node j $e_{ij} =< v_i, v_j >$ will be added in E.



Claim: Source post:
"NEWS: **'Emergency Emergency'** was the **final distress call** that was sent from the cockpit of flight **#4U9525.**

Looks terribly off course on a map. Nothing earlier? Like much earlier?

it's a scale. Emergency < Mayday.

is that a normal call in this situation?

I doubt any pilot would not say 'Emergency', but rather 'Mayday'

No, then you would say PANPAN. Trust me, I'm a pilot! Besides: Mayday is when lives are in danger...

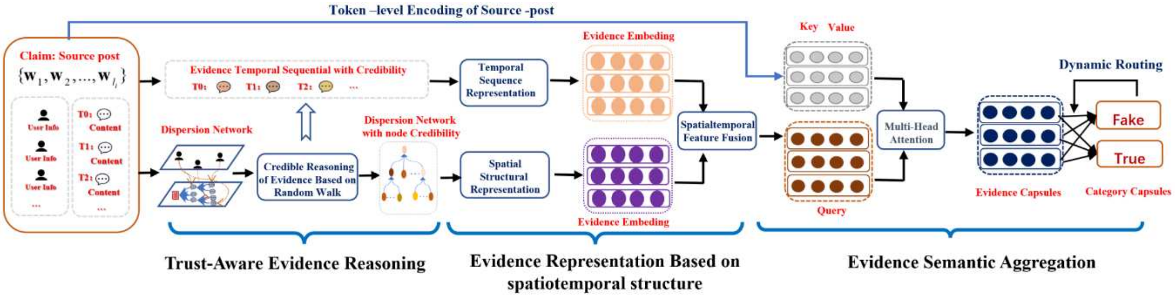
"V": false, "FR": 58408, "FL": 383, "F": 350...
"V": false "FR": 77, "FL":178, "F":1866...
"V": false "FR": 1701, "FL":1297, "F":268...
"V": false "FR": 853, "FL": 86, "F": 187 ...
"V": false "FR": 1567, "FL": 1194, "F": 1180 ...
"V": false "FR": 1567, "FL": 1194, "F": 1180 ...
Timelines
1
2
3
4
5
6
7

Figure 2. A piece of news on PHEME, and the meta-data of users participating in the discussion, the relevant comments or forwarded message on social media over time. The unit of time axis is minute. "V" represents whether the account is verified or not," FL", "FR", and "F" represent the numbers of followers, friends, and favorites, respectively. The bold words are the false portions of source information, and some explainable comments can directly confirm the falsehood of these words in the news**.**

The interpretable fake news detection task can be described as learning a decision function $f: f(S, P, G) \rightarrow y$ that maximizes prediction accuracy with reliable evidences and marks the false portions of source information with explainable evidence-related words.

4. TRSA: Trust-aware Evidence Reasoning and Spatiotemporal Feature Aggregation Model

In this section, we describe the details of the TRSA model. Its architecture is presented in Figure 3 and involves three modules: (1) a trust-aware evidence reasoning module for calculating the credibility of nodes in the information dispersion network based on a random walk model to discover high-quality evidence; (2) an evidence representation module with three units (temporal sequence representation unit, spatial structure representation unit, and fusion gate unit) for capturing the characteristics of high-level spatiotemporal structure and enhancing the semantic representation of evidence; and (3) an evidence semantic aggregation module for deepening the semantic interactions between evidence and source information, and modeling the implicit bias of evidence relative to source information based on a capsule network.



Token –level Encoding of Source -post
Claim: Source post
$\{w_1, w_2, \ldots, w_{l_i}\}$
User Info
User Info
User Info
...
T0: Content
T1: Content
T2: Content
...
Evidence Temporal Sequential with Credibility
T0: T1: T2: ...
Dispersion Network
Credible Reasoning of Evidence Based on Random Walk
Dispersion Network with node Credibility
Temporal Sequence Representation
Spatial Structural Representation
Evidence Embeding
Evidence Embeding
Spatialtemporal Feature Fusion
Key Value
Query
Multi-Head Attention
Dynamic Routing
Fake
True
Evidence Capsules
Category Capsules
Trust-Aware Evidence Reasoning
Evidence Representation Based on spatiotemporal structure
Evidence Semantic Aggregation

Figure 3. Framework of the proposed TRSA model.

4.1. Trust-aware evidence reasoning

Due to the openness and low barrier to entry of social media, erroneous may be released by artificial accounts during message propagation, which can introduce noise into fake news detection model. To avoid the impact of low-quality posts, we must calculate the credibility of posts and take highly credible posts as evidence to detect the truthfulness of source information. From the points of the publisher features and content of posts, two indicators are considered.

(1) Authority of users who publish comments: The higher the authority of users, the more reliable their comments [39]. In other words, users tend to receive information published by users with high authority.

(2) Degree of recognition of other comments in the information propagation process: Comments recognized by other highly credible comments have high credibility.

To comprehensively consider these indicators together, we construct a random walk model based on an information dispersion network that considers these indicators as the jump probabilities of random walkers in the network. The probability that random walkers will eventually travel to each node can be considered as the credibility of posts.

4.1.1. Information Dispersion Network Construction

Comment credibility ranking is based on an information dispersion network aimed at one claim (source information). We first construct an information network $G(\boldsymbol{S}) =< V, E >$ based on source information along with its related comments, as well as the authority of users participating in the discussion, as shown in Figure 4.

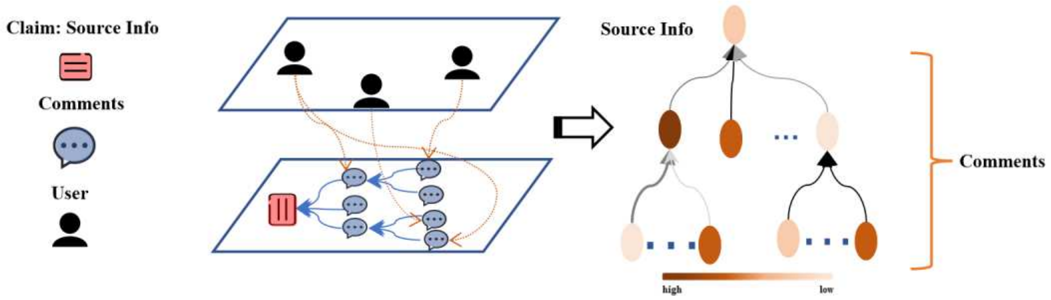


Figure 4. Information dispersion network (node color represents node credibility and edge color represents the degree of recognition of content between nodes).

The node at the top level of the network represents the source information and the other nodes represent comments. As described in Section **3,** each node is represented as $v_j = (uj, cj) \in V$, where $u_j \in R$ represents the authority of the user who posted comment j,$c_j \in R^d$ is a d-dimensional vector obtained by a pre-trained BERT model that can be used to characterize the comment content. The initial weight of node i is represented by the corresponding user's authority u_j. Each edge $e_{ij} =< v_i, v_j; w_{ij} >\in E$ represents the interaction between posts i and j. Edge weights w_{ij} indicates the recognition degree of post i relative to the content of post j. Its calculation process is as follows:

$$w_{ij} = \text{sign}(\text{content}_i, \text{content}_j) * \text{simlar}(content_i, content_j), \tag{1}$$

where $sign(content_i, content_j)$ represents the emotional difference between posts. $sign(content_i, content_j) = 1$ if the emotional polarity of the two posts is same, $sign(content_i, content_j) = 0$ otherwise. $simlar(content_i, content_j)$ represents the semantic similarity, its value in [0, 1]. We adopt the interface provided by the Baidu AI platform for calculating emotional difference[1] and adopt soft cosine measure [40] between embeddings of comments (or the source post and its comments) as semantic similarity. From Equation (1), although there is interaction between posts i and j (e.g., i is a comment to j), w_{ij} may stll be 0. This is because there is no emotional

[1] https://aip.baidubce.com/rpc/2.0/nlp/v1/sentiment_classify

resonance which means $sign(content_i, content_j) = 0$ or the content of the two posts is irrelevant which means $simlar(content_i, content_j) = 0$.

4.1.2. Credible Reasoning of Evidence Based on a Random Walk

Based on the information dispersion network, random walkers can walk randomly in the network in two ways: jumping randomly according to the weights of network nodes (i.e., considering the authority of users who publish posts) and walking randomly along to the edges in the network (i.e., considering the information interactions in the dispersion process).

The probability p_{ij} of random walkers jumping from node i to node j according to the node weight is defined as follows:

$$p_{ij} = \frac{exp(u_j)}{\sum_i exp(u_i)}, \tag{2}$$

where $\sum_n exp(u_n)$ represents the summation of weights of all nodes. According to Equation (2), the probability of random jump is only correlated with the goal node weight and its value is in (0,1). Let $\boldsymbol{P}$ denots a jump matrix, the elements in each column are the same and the sum of elements in each row is 1.

Walking according to edge weights means that random walkers select a node to reach an adjacent node directly with a certain probability along the edges of the network. If the probabilities of moving along edges are expressed by an edge transfer matrix **S**, then the probability of a random walker moving from node i to node j along an edge is expressed as follows:

$$s_{ij} = \begin{cases} 0, e_{ij} \notin E \\ w_{ij}, e_{ij} \in E \end{cases} \tag{3}$$

We let α represent the probability of random walkers walking along an edge, which is called the damping coefficient, and $1-\alpha$ is the probability of random walkers jumping according to node weight. The walking process of random walkers in the information dispersion network is described as follows:

$$\boldsymbol{r}^{(t+1)} = \boldsymbol{r}^{(t)}(\alpha\boldsymbol{S} + (1-\alpha)\boldsymbol{P}) \tag{4}$$

Here $\boldsymbol{r}^{(t)}$ and $\boldsymbol{r}^{(t+1)} \in R^{\,n}$ are n-dimensional vectors denoting the visiting probability distribution of random walkers to all nodes in the information dispersion network before and after the update, respectively. Their elements are in [0, 1]. Initially, $\boldsymbol{r}^{(0)} = (1,0, \ldots, 0)$. Langville et al. [41] pointed out that this type of random walk algorithm converges to a unique vector when the transition matrix satisfies the irreducible and periodic properties. We prove that the transfer matrix constructed in this paper satisfies this property in Appendix B. Therefore, Equation (4) eventually converges to a stable vector after multiple iterations, which can be considered as the credibility of comments.

4.2. Evidence Representation Based on Spatiotemporal Structure

As shown in Figure 2, the propagation process of source information can be expressed by a spatiotemporal structure graph. The temporal sequence can reflect the dynamic evolution of comments (evidences) content over time, while the spatial structure can reflect the real semantic interactions between evidences. To alleviate the semantic sparsity issue, we enrich the semantic representation of evidences by aggregating the temporal neighborhood and spatial neighborhood information of evidences based upon the information propagating spatiotemporal structure. Additionally, considering the difference in evidence quality, evidence credibility should be integrated into the evidence representation module to enhance the reliability of evidences semantic representation. Specifically, we consider three types of units: a temporal sequence representation unit, a spatial structure representation unit, and a fusion gate unit.

4.2.1. Evidence Temporal Sequence Representation Unit

The evolution of source information is triggered by a sequence of forwards or comments over time, as shown in Figure 2. We aim to exploit the initial semantic representation of source information and related evidence posts (comments or forwarded content) in combination with the temporal structure $P(\boldsymbol{S}) = \{(\boldsymbol{c}_0, t_0), \ldots, (\boldsymbol{c}_j, t_j), \ldots\}$ to learn evidence representation regarding temporal sequence. To obtain a more reliable sequence representation, we integrate the reliability of temporal neighborhood information into the sequence modeling process and use bidirectional long short-term memory (Bi-LSTM) [42] to model the temporal dependency of information.

$$\overrightarrow{\boldsymbol{h_\iota}} = \overrightarrow{LSTM}(exp(r_i)\boldsymbol{c_i}) \tag{5}$$

$$\overleftarrow{\boldsymbol{h_\iota}} = \overleftarrow{LSTM}(exp(r_i)\boldsymbol{c_i}) \tag{6}$$

$$\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h_\iota}}; \overleftarrow{\boldsymbol{h_\iota}}] \tag{7}$$

The above equations demonstrate how to model the temporal sequence representation of evidence i. Specifically, we apply the corresponding credibility weights (calculated in Section 4.1.2 upon each piece of posts in P and capture a reliable context from temporal neighbor posts around evidence i. In the equations, $\overrightarrow{h_\iota} \in R^land\overleftarrow{h_\iota} \in R^l$ enote the hidden state of the forward LSTM $\overrightarrow{LSTM}$ and backward LSTM $\overleftarrow{LSTM}$. l is the number of hidden units in the LSTM and [:] denotes a concatenation operation. $r_i \in [0,1]$ represents the credibility of information. To avoid the vanishing gradient problem when r_i is too small, we use $exp(r_i)$ instead of r_i. $\boldsymbol{h_i} \in R^{2l}$ is a semantic representation of the temporal structure of evidence i (when i = 0, it represents source information).

4.2.2. Evidence Spatial Structure Representation Unit

To capture the local spatial structure characteristics of evidence posts interactions in message propagation, inspired by graph attention networks (GATs) [43], we adopt attention mechanisms on the information dispersion network. Although GATs can automatically capture the contributions (attention weights) of different nodes to the target node when aggregating neighboring nodes, it ignores the credibility of neighborhood information. Therefore, such an aggregation may result in excessive noise. We introduce the credibility of posts in the process of spatial neighborhood aggregation to enhance the reliable representation of evidences, which can be formulated as

$$\beta_{ij} = \boldsymbol{softmax}(Leaky\ Re\ L\ U(\boldsymbol{\alpha}^T[exp(r_i)\boldsymbol{c}_i; exp(r_j)\boldsymbol{c}_j])), \tag{8}$$

$$\boldsymbol{m_i} = \overset{K}{\underset{k=1}{||}} \sigma(\sum_{j \in N_i} \beta_{ij}^k \boldsymbol{W}^k \boldsymbol{c}_j). \tag{9}$$

Equation (8) demonstrates how to model the semantic contribution of neighborhood post to target post (evidence). Specifically, we use a feed forward network as attention function which contains a single hidden layer with a LeakyReLU and use the global credibility weights to optimize the local semantic contribution of information. In the equation, $\boldsymbol{\alpha} \in R^{2d}$ is a learnable parameter weight vector and β_{ij} denotes semantic contribution of neighborhood node j to target node i. Equation (9) demonstrates how to aggregate spatial neighborhood information according to the calculated contributions. The spatial neighborhood representation of an evidence is obtained by a weighted summation over all spatial neighborhood semantic representations. To capture diversified representations of spatial structure relationships, attention is expanded to multi-head attention. Specifically, Equation (8)(i.e., the attention operation) is repeated K times, then the learned representations are concatenated. | | denotes a concatenation operation, $\sigma(\cdot)$denotes the exponential linear activation function, N_i denotes the collection of posts directly connected with evidence i in the information dispersion network, $\boldsymbol{W}^k \in R^{q \times 2d}$ is a learnable shared parameter matrix that acts on each node in the network, and $\boldsymbol{m_i} \in R^{Kq}$ is the semantic representation of the spatial structure of evidence i (when i = 0, it represents source information).

4.2.3. Spatiotemporal Feature Fusion Unit

To represent the semantic features of information from multiple perspectives, the temporal semantic representation and spatial structural representation of evidence are selected and combined through a fusion gate to obtain a semantic representation of the spatiotemporal structure of evidence. Because the temporal semantic representation and spatial structural semantic representation of evidence are not in the same semantic space ($2l \neq Kq$), it is necessary to convert them into the same semantic space (i.e., $\boldsymbol{h}', \boldsymbol{m}' \in R^{\,h}$).

$$\begin{aligned} \boldsymbol{h}_i' &= tanh(\boldsymbol{W}_h \boldsymbol{h}_i) \\ \boldsymbol{m}_i' &= tanh(\boldsymbol{W}_m \boldsymbol{m}_i) \end{aligned} \tag{10}$$

Here, $\boldsymbol{W}_h \in R^{\,h\times 2l}$ and $\boldsymbol{W}_m \in R^{\,h\times Kq}$ denote the transformation matrixes from the different feature spaces of evidence i to an implied common space.

Next, we use a fully connected layer with a sigmoid activation function to learn the importance of temporal sequence semantic representation and spatial structure semantic representation, as shown in Equation (11):

$$\boldsymbol{z} = \sigma(\boldsymbol{W}_z[\boldsymbol{h}_i'; \boldsymbol{m}_i']), \tag{11}$$

where $\boldsymbol{W}_z \in R^{h\times 2h}$ is a learnable weight matrix. $\boldsymbol{z} \in R^{\,h}$ is a weight vector to trade-off the semantic reprensentation from spatial and temporal structure. Its elements are in (0,1). Finally, the two semantic representations are fused using different weights.

$$\boldsymbol{x}_i = \boldsymbol{z} \odot \boldsymbol{h}_i' + (\boldsymbol{1} - \boldsymbol{z}) \odot \boldsymbol{m}_i' \tag{12}$$

Here, $\odot$ denotes elementwise multiplication. $\boldsymbol{x}_i \in R^h$ is the information representation obtained through the fusion gate, and h is the dimension of the fusion gate output representation.

4.3. Evidence Semantic Aggregation Based on a Capsule Network

Based on the spatiotemporal semantic representation of evidence, fake information can be detected by aggregating the implicit bias of evidence to evaluate the truthfulness of source information. We incorporate a capsule network [44] into our model to model the implicit bias of evidence toward claims. This process is illustrated in Figure 3, where we first model the semantic interactions between evidence and source information to capture the controversial points (false portions) of source information and form evidence capsules (low-level capsule). We then aggregate the implicit bias of each evidence capsule regarding the source information through a dynamic routing mechanism.

4.3.1. Semantic Interactions between Evidence and Source Information Based on Multi-head Attention

Although the fusion gate can efficiently aggregate the spatiotemporal neighborhood information of evidences to obtain a reliable evidence representation, it cannot model the fine-grained semantic interactions between evidence and source information. To capture the focus of evidences on source information, we adopt a multi-head attention mechanism [46] to model the semantic interactions between evidence and source information. Specifically, we consider the evidence representation set $\boldsymbol{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_m\} \in R^{\,(m+1)\times h}$ obtained in Section 4.2 as query vectors Q and consider the semantic representation of source information $\boldsymbol{S} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_l\} \in R^{l\times d}$as keys (K) and values (V). We use each piece of evidence in $\boldsymbol{X}$ to assign attention to each word in $\boldsymbol{S}$ through scaled dot-product attention and then apply the resulting attention weights to the source information as follows:

$$Attention(Q, K, V) = \boldsymbol{softmax}(\left[\frac{XK^T}{\sqrt{d}}\right] V) \tag{13}$$

To prevent the model from focusing too heavily on a particular location, we first map queries, keys, and values to different spaces through different types of linear transformations. We then perform attention calculations in different spaces in parallel to obtain representations of each comment (evidence) in different subspaces.

$$\begin{aligned} Head_i &= Attention(\boldsymbol{XW}_i^Q, \boldsymbol{SW}_i^K, \boldsymbol{SW}_i^V) \\ \boldsymbol{E} &= MultiHeadAttention(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{S}) \\ &= Re\,L\,U(\left[Head_1\middle|\middle|Head_2\middle|\middle|Head_3 \ldots Head_n\right]) \end{aligned} \tag{14}$$

Here, $\boldsymbol{W}_i^Q \in R^{\,h\times p}$, $W_i^k, W_i^v \in R^{\,d\times p}$and $\boldsymbol{E} = \{\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\} \in R^{\,(m+1)\times np}$ represent collections of underlying evidence capsules.

4.3.2. Evidence Aggregation Based on a Dynamic Routing Mechanism

In the fake news detection task, high-level capsules are regarded as the representations of news (source information) authenticity, namely category capsules. Specially, there are two types of category capsules, namely fake or true, in our capsule network. Each category capsule is assembled from the underlying evidence capsules by a weighted summation over all corresponding vectors. It can be described as follows:

$$\boldsymbol{v}_j = squash(\sum_{i=0}^{m} O_{j|i} \boldsymbol{W}_{ji} \boldsymbol{e}_i), j \in (0,1), \tag{15}$$

where $\boldsymbol{v}_j \in R^{\,d_v}$ is a category capsule. $O_{j|i}$is the probability that evidence e_i supports that source information belongs to category j, which can be calculated by dynamic routing mechanism on original logits b_{ji}. The specifics of this process are provided in Algorithm 1. $\boldsymbol{W}_{ji} \in R^{\,d_v\times np}$ is a learned parameter matrix. To enable the module of the category capsule to determine the probability that information belongs to this category and increase nonlinear characteristics, a squash operation is applied to compress the module length of the capsule to [0,1].

$$\boldsymbol{v}_j = squash(\boldsymbol{v}_j) = \frac{\|\boldsymbol{v}_j\|^2}{1 + \|\boldsymbol{v}_j\|^2} \frac{\boldsymbol{v}_j}{\|\boldsymbol{v}_j\|} \tag{16}$$

Algorithm 1 Dynamic Routing Mechanism

```
Input: $\boldsymbol{W}_{ji}, \boldsymbol{e}_i$
Output: $\boldsymbol{v}_j$
1: Init the coupling parameter $b_{ji} == 0$
2: for each iteration do
3:      Update $O_{j|i} = soft\,max(b_{ji})$
4:      Update all the class capsules based on Equation (15)
5:      Update $b_{ji} = \boldsymbol{W}_{ji}\boldsymbol{e}_i \cdot \boldsymbol{v}_j$
6: end for
7: return $\boldsymbol{v}_j$
```

4.3.3. Detection

After category capsules are obtained through the dynamic routing mechanism, the category capsule with the largest module length is chosen as the representation of news (source information) truthfulness.

$$\hat{y} = max(\|v_0\|, \|v_1\|) \tag{17}$$

Finally, the cross-entropy loss is used to capture the error between forecast results and factual value:

$$L(\boldsymbol{\theta}) = -\sum_i y_i\, log(\hat{y}_i), \tag{18}$$

where $\boldsymbol{\theta}$ denotes the model parametric set and $y_i \in \{0,1\}$ is the ground-truth label of the i-th instance

5. Experiments

In this section, we present experiments conducted on public datasets to evaluate the effectiveness of the TRSA model. Particularly, we aim at answering the four evaluation issues as follows:

- **EI1:** Can TRSA achieve better performance than the state-of-the-art models?
- **EI2:** How effective is each component of TRSA at improving detection performance?
- **EI3:** Can TRSA make detection results easy to understand using the evidence reasoning and evidence aggregation module?
- **EI4:** What is the performance of the model for the early detection of fake news?

5.1. Experimental Datasets and Settings

5.1.1. Datasets

We evaluate our model on two real datasets: PHEME (English dataset, mainly from foreign Twitter) [46] and CED (Chinese dataset, mainly from the domestic Sina platform)[47]. The PHEME dataset contains three types of labels: true, fake, and uncertain. The CED dataset contains only true and fake labels. Because CED lacks the basic information of the users participating in a discussion, we collected basic information on the users participating in discussions by designing a web crawler[2]. Table 1 provides the detailed statistics of datasets.

Table 1. Statistics of datasets.

Statistical Indicators	PHEME	CED
Source Tweets	2402	3387
Comments/rep	30,723	1,275,179
Users	20,538	1,064,970
Fake	638	1538
True	1067	1849
Uncertain	697	-

5.1.2. Comparison Methods

We compare TRSA to the following baselines:

- **DTC [5]:** This method utilizes multi-dimensional statistical features from the four perspectives of text content, user characteristics, forwarding behavior, and communication mode, and implements decision trees to determine the truthfulness of information.
- **SVM-TS [48]:** This method utilizes SVMs with linear kernel function to model temporal features for false information.
- **HSA-BLSTM [49]:** HSA-BLSTM is a hierarchical neural network model used to describe the semantic features of different levels of rumor events (a rumor event is composed of source information and multiple forwarded or commented posts, and each post is composed of words).
- **DTCA [18]:** This model considers user comments as an evidence source for truthfulness judgment of a claim and uses a co-attention network to enhance the semantic interactions between evidence and source information.

[2] Since some of the participating accounts have been cancelled, we only collected 9 types of meta-features of about 460 thousand related accounts, including gender, location, description, message, followers, friends, etc. The values of the cancelled accounts' multiple meta-features are given 0.

- **BERT-Emo [35]:** BERT-Emo uses a pretrained language model to obtain the text semantic representation and the emotions difference between an information publisher and their audience.
- **GLAN [19]:** GLAN is a novel neural network model that can corporately model local semantic features and global propagating features.
- **BiGCN [20]:** BiGCN is a two-layer graph convolutional network model to capture the bidirectional propagating structure of information. It also integrates source post information into each layer of the GCN to enhance the impact of source information.
- **DDGCN [28]:** DDGCN is a dynamic graph convolution neural network model to capture the characteristics of the information propagation structure and knowledge entity structure at each point in time. Since our model only concentrates on the contents and social contexts, we don't introduce dynamic knowledge structure.

5.1.3. Experiment Setup

The environmental configurations of all experiments in this study are as follows: Intel Core i9 CPU, 64 GB of RAM, GTX-3090 GPU.

The experimental environments and parameters of all compared methods in this study are set according to the original reports. We use the PyTorch framework to implement our models. The model parameters are optimized and updated using the Adam optimizer. In our model, we use pre-trained BERT models (bert-base-uncased for English and bert-base-Chinese for Chinese) to initialize the vector representations of text information. We employ accuracy(A), precision(P), recall(R), and F1 as assessment indicators. Model parameter details are provided in Appendix C.

5.2. Performance Comparison

To answer **EI1**, we contrast TRSA with baseline models on two real datasets. The experimental results are reported in Table 2. The bold values represent the best results, and the underlined values represent the second-best results.

Table 2. Results contrast between different methods.

Methods	PHEME				CED			
	A	P	R	F	A	P	R	F1
DTC	0.669	0.678	0.678	0.667	0.731	0.731	0.719	0.725
SVM-TS	0.722	0.788	0.758	0.721	0.857	0.859	0.858	0.859
HSA_BLSTM	0.757	0.772	0.731	0.745	0.878	0.877	0.876	0.876
DTCA	0.823	0.861	0.791	0.825	0.901	0.921	0.891	0.902
BERT-Emo	0.800	0.795	0.795	0.793	0.905	0.916	0.913	0.914
GLAN	0.828	0.824	0.822	0.823	0.918	0.917	0.914	0.915
BiGCN	0.847	0.840	0.834	0.835	0.919	0.918	0.916	0.917
DDGCN	0.855	0.846	0.841	0.844	0.922	0.920	0.931	0.925
TRSA	**0.885**	**0.896**	**0.871**	**0.881**	**0.953**	**0.950**	**0.954**	**0.952**

We can obtain several observations as follows:

- The deep neural network models are superior to the models based on feature engineering (DTC, SVM-TS). The most fundamental reason is that deep neural network models can automatically learn implicit high-level semantic representations, whereas traditional machine learning methods that rely on feature engineering can only capture obvious false information in the presentation layer, which leads to various limitations.
- The models that add semantic interactions between claims and comments (DTCA, BERT-Emo) perform better than the model that work with text and hierarchical time-series structure (HSA_BLSTM). DTCA automatically captures controversial portions of source information through a co-attention mechanism. BERT-Emo model constructs a dual emotional feature set by

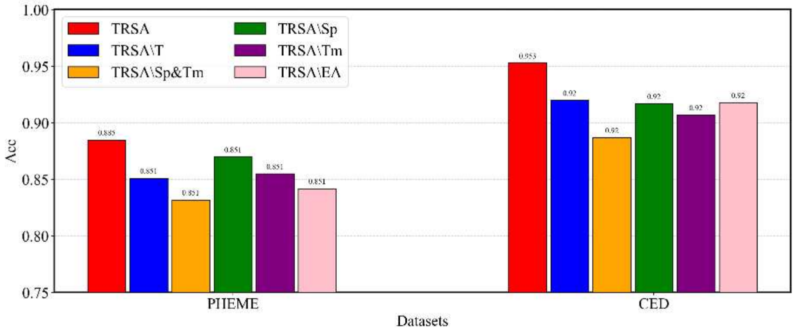
measuring the difference between the emotions of an information publisher and their audience to improve false information detection performance.

- The models based on information propagation structure are superior to the models based on text semantics (DTCA, BERT-Emo, HAS-BLSTM). For example, GLAN, BiGCN, and DDGCN achieve improvements of approximately 0.5% to 3.2% in terms of accuracy on the two datasets compared to DTCA. This indicates that mining the hidden structural features of information propagation is very helpful for improving detection performance. However, in terms of precision, because DTCA uses decision trees to filter out some low-credibility noise comments, its performance is approximately 1.5% higher than that of the aforementioned models on PHEME. Moreover, it can be observed that DDGCN shows better performance than BiGCN and GLAN, indicating that spatiotemporal structure features can finely depict the semantic interaction in message propagation and thus improve performance.
- The proposed model outperforms most post-based models and propagation-based models in terms of most indicators on the two real datasets. Compared to DTCA, the proposed model enriches claim and comment semantic information from the perspective of time and space propagation structures. Its performance is 5.7%, 3.2%, 7.15%, and 5.3% higher than that of DTCA in terms of accuracy, precision, recall, and F1, respectively. Compared to DDGCN, these four indicators are 3%, 4%, 2.65%, and 3.5% higher on average. This is because DDGCN treat all comments equally, which introduces noise. In contrast, our model reduces noise by calculating the credibility of comments.

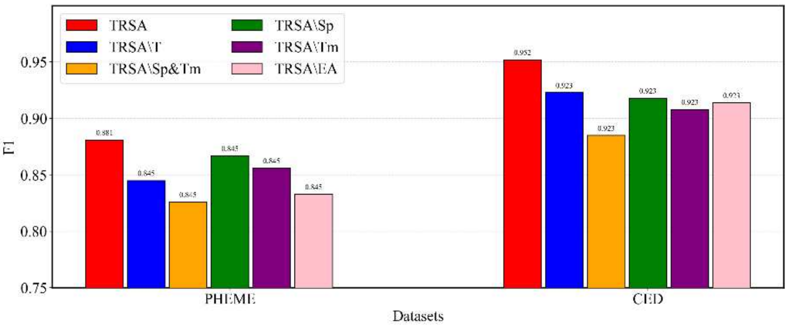
5.3. Ablation Study

To answer **EI2**, we investigate how effective are the key components on TRSA by designing five variations: (1) **TRSA\T** removes the trust-aware evidence reasoning module. (2) **TRSA\Sp** removes the spatial characteristics of information propagation. (3) **TRSA\Tm** removes the temporal characteristics of information propagation. (4) **TRSA\Sp&Tm** removes the spatiotemporal characteristics of information propagation. (5) **TRSA\ EA** replaces evidence aggregation with a max-pooling layer and a fully connected layer.

In Figure 5, one can see that all variations performed less well than the complete TRSA model on both datasets. Specifically, when removing the spatiotemporal characteristics of information propagation, the F1 drops by 5.5% on the PHEME dataset and 6.7% on the CED dataset. This indicates the necessity of the temporal and spatial propagation structure information to improve model performance. Furthermore, the results demonstrate that removing spatial structure make a larger decrease in model performance compared to removing the temporal structure. This indicates that the spatial structure is more effective than the temporal structure. When removing the trust-aware evidence reasoning module, the decrease in terms of F1 on PHEME is 3.6% and that on CED is 2.9%. This demonstrates that the impact of low-quality comments on the performance of the model can be mitigated by the evidence credibility index. The replacement of the evidence aggregation module led to a decrease in F1 of 4.8% on PHEME and 3.8% on CED. This demonstrates the necessity of aggregating evidence semantics to achieve better performance.



(**a**) Accuracy contrast among ablation variations on the PHEME and CED datasets



(**b**) F1 score contrast among ablation variations on the PHEME and CED datasets

Figure 5. Results contrast among ablation variations on the PHEME and CED datasets.

5.4. Explainable Analysis

The trust-aware evidence reasoning and evidence aggregation modules make the decision-making process of TRSA more transparent and the results more interpretable. To answer **EI3**, we visualize the evidence credibility, attention weight distribution, and implicit biases of evidence when predicting fake news. Figure 6 presents the results for a specific sample in the PHEME testing set.

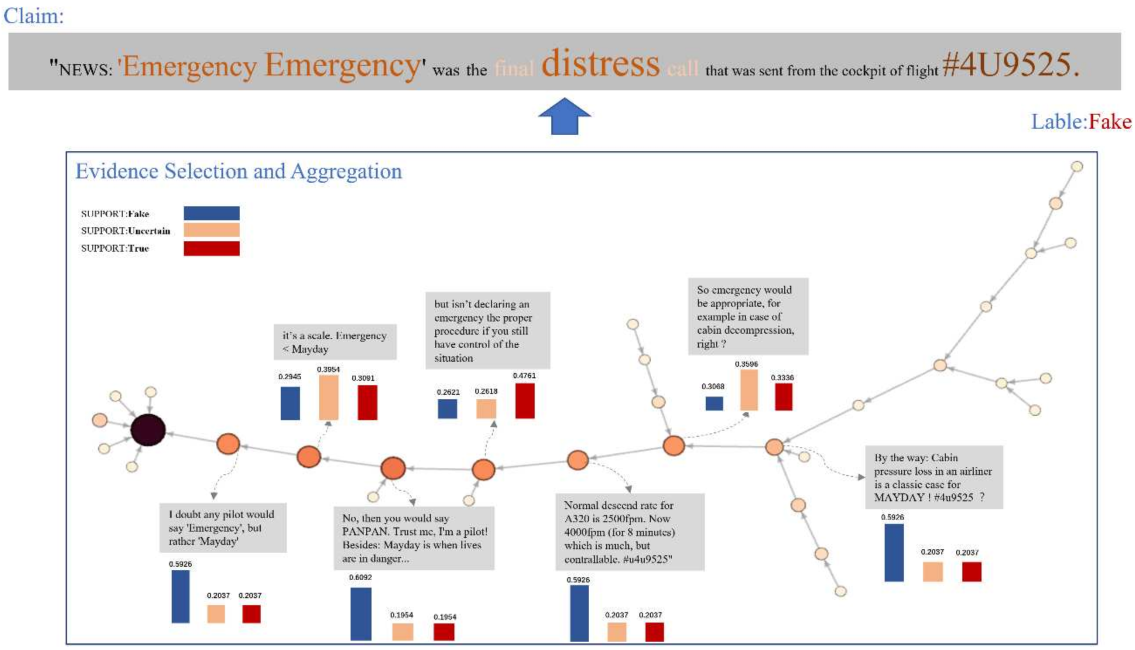

Claim:
"NEWS: 'Emergency Emergency' was the final distress call that was sent from the cockpit of flight #4U9525.
Lable:Fake
Evidence Selection and Aggregation
SUPPORT:Fake
SUPPORT:Uncertain
SUPPORT:True
it's a scale. Emergency < Mayday
0.2945
0.3954
0.3091
but isn't declaring an emergency the proper procedure if you still have control of the situation
0.2621
0.2618
0.4761
So emergency would be appropriate, for example in case of cabin decompression, right ?
0.3068
0.3596
0.3336
I doubt any pilot would say 'Emergency', but rather 'Mayday'
0.5926
0.2037
0.2037
No, then you would say PANPAN. Trust me, I'm a pilot! Besides: Mayday is when lives are in danger...
0.6092
0.1954
0.1954
Normal descend rate for A320 is 2500fpm. Now 4000fpm (for 8 minutes) which is much, but contrallable. #u4u9525"
0.5926
0.2037
0.2037
By the way: Cabin pressure loss in an airliner is a classic case for MAYDAY ! #4u9525 ?
0.5926
0.2037
0.2037

Figure 6. A case study on the explainability of TRSA by visualizing the detection process of one example (labeled fake).

- First, we focus on each token in the source information by accumulating the attention values of the interactions between evidence (high-quality comments) and claims (source information) in the information propagation process, which is represented by the size and color of each word. The larger the font, the darker the color of the word, indicating that more attention is assigned to the word in the process of information propagation and that the word is more controversial. One can see that "*Emergency*", "*distress*", and "*# 4U9525*" have been widely discussed by users in the process of information propagation, which further demonstrates that our model can automatically capture controversial content.
- Second, we use Gephi to draw the information dispersion network, where the sizes of nodes are determined by their credibility (the higher the credibility of the node, the larger the node). One can see that the black nodes represent source information, and the other nodes represent related forwarding or comment posts. Comments endowed with high credibility weights can be used

as evidence to prove that the source information is fake. Consider the following comments. "*I doubt that any pilot would not say 'Emergency,' but rather'Mayday'.*" "*No, then you would say 'PANPAN'. Trust me, I'm a pilot! Besides, 'Mayday' is a time when life is in danger.*" "*By the way: Cabin pressure loss in an airliner is a classic case for Mayday! \# 4u9525?.*". The "*PANPAN*" and "*Mayday*" terms appearing in these comments are internationally used radio crisis call signals indicating that the "Emergency" term in the source information is incorrect. This indicates that the trust-aware evidence reasoning module can provide highly reliable evidence to explain the model results. To measure the support of evidence for results objectively, we examined the implicit bias distribution of evidence by visualizing the aggregation probabilities of the underlying evidence capsules into the high-level category capsule in the evidence aggregation module. One can see that most of highly credible evidences refutes the source information content.

- To unfold user attention distribution differences between fake and true news content, we randomly select three fake (0–2) and three true (3–5) news stories, and plot their token weights distribution based on the attention of the interactions between the evidence and claims. As shown in Figure 7, the horizontal direction from left to right represents the word sequence. In the vertical direction, the first three entries represent fake information (0–2) and the last three represent true information (3–5). One can see that some local fake news content attracts widespread attention, whereas the attention on each component of the real news is relatively uniform.

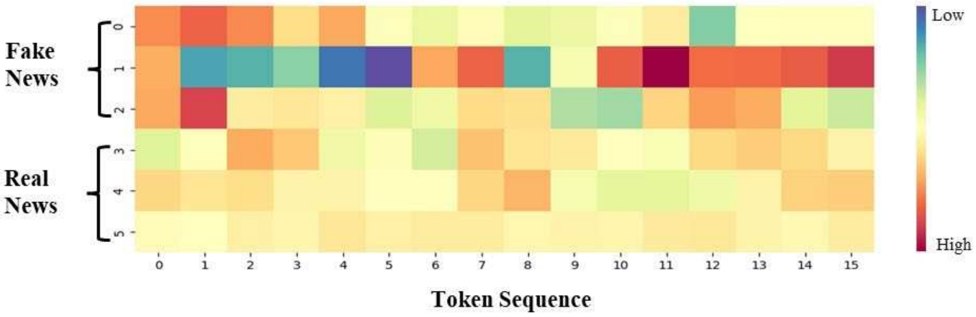
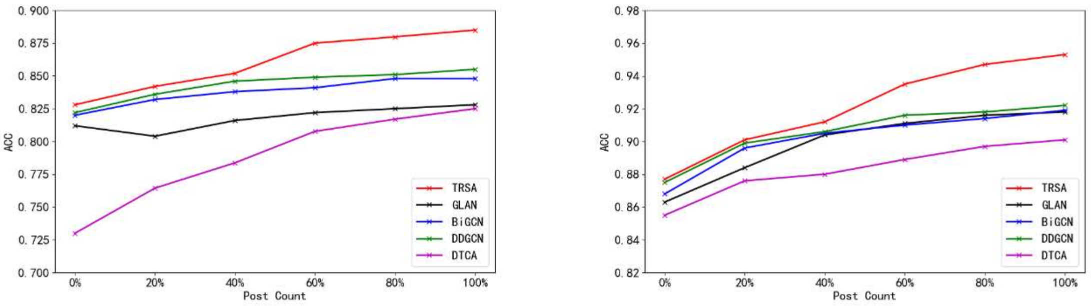


Figure 7. Visualization of word-level attention weights in propagations of three fake (0–2) and three true (3–5) source tweets.

5.5. Early Fake News Detection Performance

To answer **EI4**, we sorted all comments (or forwarded posts) according to their publishing time and evaluated the change in TRSA's detection performance by changing the number of comments received (0%, 20%, 40%, 60%, 80%, 100%). Figure 8 presents the early detection results of the model for both datasets. One can see that when there only the first 40% of comments are considered, the accuracy of the proposed model can reach 85.2% and 91.2% on the two datasets, which is superior to the results of the baseline models. This indicates that our model performs well in terms of early detection. Additionally, we observed that the accuracies of the GLAN, BiGCN, and DDGCN models increase slowly over time, whereas the proposed model exhibit significantly improved performance over time. This is because the dispersion network structure of information becomes more complex and the types of posts become more diversified over time. The proposed model has module for filtering noise posts. Therefore, they have good robustness.



(**a**) PHEME Dataset (**b**) CED Dataset

Figure 8. Fake news early detection performance contrast between different models.

6. Conclusions

We propose an interpretable fake news detection method called TRSA based on trust-aware evidence reasoning and spatiotemporal feature aggregation. First, from the perspective of user authority and dispersion network structure, the trust-aware random walk concept is used to calculate the credibility of posts during the process of information dissemination. Second, by considering the credibility of nodes, the semantic information of nodes are aggregated from the perspective of temporal and spatial structure to obtain a reliable representation of evidence. Finally, to increase the interpretability of our model, we capture the controversial points of source information based on multiple attention and model the implicit bias of an evidence capsule through a dynamic routing mechanism to enhance the transparency of model reasoning. Experimental results indicate that our model is effective and interpretable. In the future, we plan to optimize our work from two perspectives: (1) introducing background knowledge to enrich model explainability further and (2) integrating additional modal information such as images and videos to optimize model performance further.

Author Contributions: Conceptualization, J.C. and G.Z.; methodology, J.C. and C.L; software, C.J.; validation, J.C., Y.W. and H.L.; writing—original draft preparation, J.C.; writing—review and editing, G.Z. and C.L; visualization, J.C; funding acquisition, C.J. and G.Z.

Funding: This research was funded by Henan Province Science and Technology Project (Grant No. 222102210081).

Data Availability Statement: We evaluated our model on two public datasets: PHEME (https://figshare.com/articles/dataset/PHEME_rumour_scheme_dataset_journalism_use_case/2068650) and CED (https://github.com/thunlp/Chinese_Rumor_Dataset).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. User Authority Calculation Method Based on Multidimensional Attribute Weighted Fusion

To represent authority, principal component analysis (PCA) is used to fuse multiple metadata features of users as

$$w_i = \lambda_1 VF_i + \lambda_2 FL_i + \lambda_3 FR_i + \lambda_4 D_i + \lambda_5 GEO_i + \lambda_6 F_i \tag{19}$$

where λ_i represents the weight coefficient of the user's ith meta-feature. VF_i, D_i , and GEO_i represent whether the elements of "verified," "geo," and "homepage introduction" exist, respectively. FL_i, FR_i , and F_i represent the numbers of followers, friends, and favorites, respectively.

Table A1 lists the metadata and other information of users participating in discussion. For Boolean features, when the value is true, we convert it to a value of one and when the value is false, we convert it to a value of zero. Because the value ranges of each feature are very different, we applied the min-max normalization method to make the values of the metadata features of the six-dimensional users dimensionless while keeping their feature distribution characteristics unchanged.

Table A1. Metadata characteristics of users participating in discussion.

Data Type	Multidimensional Metadata	Weights	
		PHEME	CED
BOOL	verified(V)	1.20e-06	2.19e-07
	whether there is homepage introduction(D)	1.00e-05	2.25e-04
	whether allows the geo-spatial positioning(GEO)	1.26e-05	8.08e-06
Long Int	fans(FL)	2.11e-01	1.26e-01
	friends(FR)	9.58e-01	1.06e-02
	favorites(F)(PHEME)/message(M)(CED)	1.91e-01	9.91e-01

The PCA method was adopted to convert user six-dimensional metadata features into multiple comprehensive indicators to calculate user authority while minimizing the loss of metadata information. Figure A1 presents the relationship between the number of principal components and variance contribution rate.

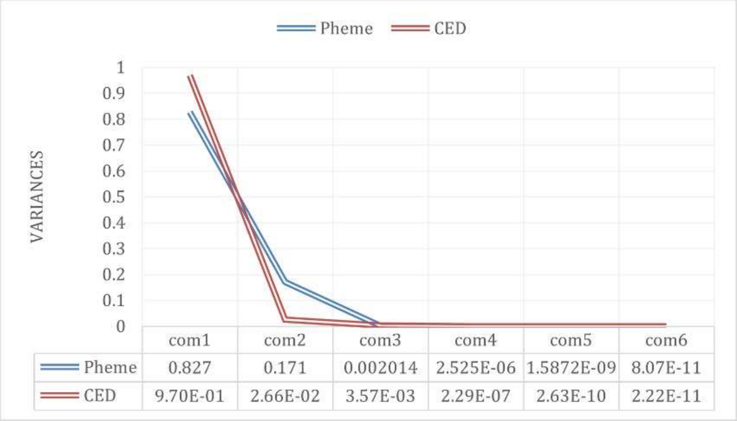


Figure A1. Variance contribution rate of different principal components on two data sets.

One can see that the variance contribution rate of the first principal component on both datasets exceeds 0.8. Therefore, to simplify our calculations, we directly represented the authority of users involved in a discussion. For the two datasets, the weights of the six-dimensional features in the first principal component are presented in the third column of Table A1.

Appendix B. A Proof of Irreducible and Aperiodic Property of Transfer Matrix

let $\mathbf{B} = \alpha\mathbf{S} + (1-\alpha)\mathbf{P}$.

First, we prove that matrix B is irreducible: Since all elements in S are greater than or equal to 0 and all elements in P are greater than 0, all elements in B are greater than 0. Therefore, the directed graph G(B) corresponding to matrix B must be strongly connected. According to **Theorem 1**: *Complex matrix B of order n (n>1) is irreducible if and only if the directed graph G(B) corresponding to matrix B is strongly connected.*, matrix B is irreducible.

Then, we prove that matrix B has aperiodic property: According to the previous analysis, the elements on the diagonal of matrix B are all greater than 0, so there are self-cyclic edges in its corresponding strongly connected graph. Therefore, matrix B also has aperiodic property.

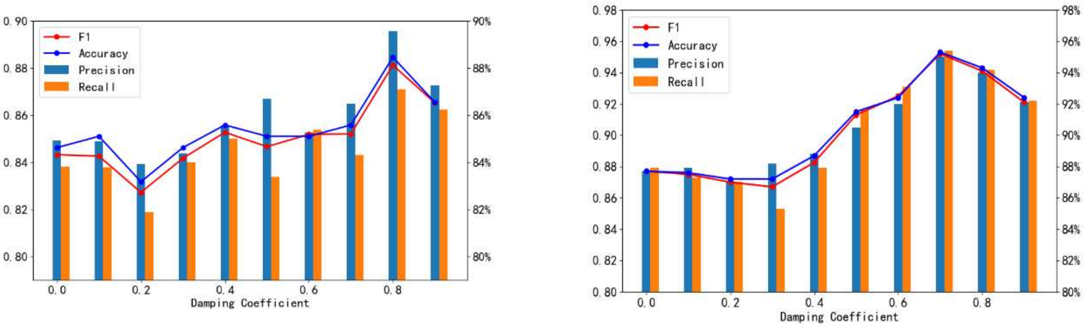
Appendix C. Optimal Parameter Configuration of the TRSA Model on Two Datasets

Table C1 presents the optimal configuration of the proposed model for the two datasets.

Table C1. Detailed configuration of model parameters.

Parameter Type	Parameter	PHEME/CED
Configuration Parameter	LEARNING_RATE	2e-5
	BATCH_SIZE	8
	MAX_SEQUENCE_LENGT	70
	LEN_COM	50
	EPOCH	15
	NHEADS	8
Hidden Parameter Configuration	LSTM_hiden size	384
	GAT_hiden size	96
	MultiHeadAttention_out size	200
	Capsule_out_dim	200

Further, we analyzed the changes in model performance under different damping coefficients α (described in Section 4.1.2 and determined the optimal damping parameter. The value range of the damping coefficient is [0,1). For $\alpha = 0$, random walkers only jump according to user authority, regardless of the actual dispersion network (i.e., the credibility of the obtained evidence is only determined by user own authority). For $\alpha \rightarrow 1$, random walkers largely ignore user authority and jump along the actual dispersion network. In Figure C1, one can see that on the PHEME dataset, when $\alpha = 0.8$ is used, the model performance is optimal and on the CED dataset, the optimal value of the damping coefficient is 0.7.



(**a**) PHEME Dataset

(**b**) CED Dataset

Figure C1. Performance analysis with different damping coefficients.

References

1. Sheng, Q.; Cao, J.; Bernard, H.R.; Shu, K.; Li, J.; Liu, H., Characterizing multi-domain false news and underlying user effects on chinese weibo. *Inf. Process. Manage.* **2022**, 59, 18.
2. Fourney, A.; Racz, M.Z.; Ranade, G.; Mobius, M.; Horvitz, E., Geographic and temporal trends in fake news consumption during the 2016 us presidential election. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Association for Computing Machinery: Singapore, Singapore, 2017; pp 2071–2074.
3. Islam MS, Sarkar T, Khan SH, et al. COVID-19-Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis. The American Journal of Tropical Medicine and Hygiene. 2020 Oct;103(4):1621-1629.
4. Bazmi, P.; Asadpour, M.; Shakery, A., Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Inf. Process. Manage*. **2023**, 60, 17.
5. Qazvinian, V., Rosengren, E., Radev, D., Mei, Q. Rumor has it: Identifying misinformation in microblogs. In Proceedings of the 2011 conference on empirical methods in natural language processing, Edinburgh, Scotland, UK, 2011; pp. 1589–1599.
6. Castillo, C., Mendoza, M., Poblete, B. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011; pp. 675–684.

7. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B. A stylometric inquiry into hyperpartisan and fake news. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia,2018; pp. 231–240.
8. Ito, J.; Song, J.; Toda, H.; Koike, Y.; Oyama, S., Assessment of tweet credibility with lda features. In Proceedings of the 24th International Conference on World Wide Web, Association for Computing Machinery: Florence, Italy, 2015; pp 953–958.
9. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of 2013 IEEE 13th International Conference on Data Mining, IEEE, 2013; pp. 1103–1108.
10. Hu, X., Tang, J., Gao, H., Liu, H. Social spammer detection with sentiment information. In Proceedings of 2014 IEEE 14th international conference on data mining, IEEE,2014; pp. 180–189.
11. Karimi, H., Tang, J. Learning hierarchical discourse-level structure for fake news detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019; pp. 3432–3442.
12. Shu, K., Wang, S., Liu, H. Beyond news contents: The role of social context for fake news detection. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, Association for Computing Machinery: Melbourne VIC, Australia, 2019; pp 312–320.
13. Wang, Y., Qian, S., Hu, J., Fang, Q., Xu, C. Fake news detection via knowledge-driven multimodal graph convolutional networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Association for Computing Machinery: Dublin, Ireland, 2020; pp 540–547.
14. Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., Zhou, M. Compare to the knowledge: Graph neural fake news detection with external knowledge. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021; pp. 754–763.
15. Yan, R., Yen, I.E., Li, C.T., Zhao, S., Hu, X. Tackling the achilles heel of social networks: Influence propagation-based language model smoothing. In Proceedings of the 24th International Conference on World wide web, Florence, Italy, 2015; pp. 1318–1328.
16. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H. defend: Explainable fake news detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery: Anchorage, AK, USA, 2019; pp 395–405.
17. Wu, L., Rao, Y., Jin, H., Nazir, A., Sun, L. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China,2019; pp. 4644–4653.
18. Wu, L., Rao, Y., Zhao, Y., Liang, H., Nazir, A. Dtca: Decision tree-based co-attention networks for explainable claim verification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020; pp. 1024–1035.
19. Yuan, C., Ma, Q., Zhou, W., Han, J., Hu, S. Jointly embedding the local and global relations of heterogeneous graph for rumor detection, in:2019 IEEE international conference on data mining (ICDM), IEEE, 2019; pp. 796–805.
20. Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., Huang, J. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 2020; pp. 549–556.
21. Wang, Y., Qian, S., Hu, J., Fang, Q., Xu, C. Fake news detection via knowledge-driven multimodal graph convolutional networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 2020;pp. 540–547.
22. Lu, Y.J., Li, C.T. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020; pp. 505–514.
23. Song, Y.Z., Chen, Y.S., Chang, Y.T., Weng, S.Y., Shuai, H.H. Adversary-aware rumor detection. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; pp. 1371–1382.

24. Huang, Z., Lv, Z., Han, X., Li, B., Lu, M., Li, D. Social bot-aware graph neural network for early rumor detection. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 2022; pp. 6680–6690.
25. Guo, B.; Jiang, Z.-b.; Kumar, M., What is the internet water army? A practical feature-based detection of large-scale fake reviews. Mob. Inf. Syst.**2023**, 2023, 11.
26. Huang, Q., Zhou, C., Wu, J., Liu, L., Wang, B. Deep spatial–temporal structure learning for rumor detection on twitter. *Neural Comput & Applic*, **2020**,1–11.
27. Xia, R.; Xuan, K.; Yu, J. In A state-independent and time-evolving network for early rumor detection in social media, Online, November, Association for Computational Linguistics: Online, 2020; pp 9042-9051.
28. Sun, M., Zhang, X., Zheng, J., Ma, G. DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media.In Proceedings of the AAAI Conference on Artificial Intelligence, 2022; pp.4611-4619.
29. Rosnow, R.L. Inside rumor: A personal journey. *American psychologist.* **1991.** 46, 484.
30. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y. Prominent features of rumor propagation in online social media. In Proceedings of 2013 IEEE 13th international conference on data mining, Dallas, TX, USA, 2013; pp. 1103–1108.
31. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, July, 2018;pp. 231–240.
32. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.-F.; Cha, M., Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, AAAI Press: New York, New York, USA, 2016; pp 3818–3824..
33. Wang, Y., Wang, L., Yang, Y., Lian, T. SemSeq4FD: Integrating global semantic relationship and local sequential order to enhance text representation for fake news detection. *Expert Systems with Applications,* **2020**, 166, 114090.
34. Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., Zhou, M., 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, August, 2021; pp. 754–763.
35. Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; Shu, K., Mining dual emotion for fake news detection. In Proceedings of the Web Conference 2021, Association for Computing Machinery: Ljubljana, Slovenia, 2021; pp 3465–3476.
36. Shi, B., Weninger, T. Discriminative predicate path mining for fact checking in knowledge graphs. *Know. -Based Syst*. **2016**, 104, 123–133.
37. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A. Computational fact checking from knowledge networks. *PloS one*, 2015,10.
38. Chen, T., Li, X., Yin, H., Zhang, J. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection, in: Paciffc-Asia conference on knowledge discovery and data mining, Cham, 2018; pp. 40–52.
39. Shan, Y., How credible are online product reviews? The effects of self-generated and system-generated cues on source credibility evaluation. *Comput. Hum. Behav*,**2016**, 55, 633–641.
40. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, **2014**,18, 491–504.
41. Langville, Amy N. and Meyer, Carl D. Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton: Princeton University Press, 2006.
42. Graves, A., Fernández, S., & Schmidhuber, J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. International Conference on Artificial Neural Networks, 2005.
43. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. Graph attention networks. In Proceedings of the 6th International Conference on Learning Representations,2018.
44. Sabour, S.; Frosst, N.; Hinton, G.E., Dynamic routing between capsules. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc.: Long Beach, California, USA, 2017; pp 3859–3869.

45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I., Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc.: Long Beach, California, USA, 2017; pp 6000–6010.
46. Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., Tolmie, P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*. **2016**,11,3.
47. Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., Sun, M. Ced: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*. **2021**,33, 3035–3047.
48. 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, Association for Computing Machinery: Melbourne, Australia, 2015; pp 1751–1754.
49. Guo, H.; Cao, J.; Zhang, Y.; Guo, J.; Li, J., Rumor detection with hierarchical social attention network. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery: Torino, Italy, 2018; pp 943–951.
50. Sun, M., Zhang, X., Zheng, J., Ma, G. DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media.In Proceedings of the AAAI Conference on Artificial Intelligence, 2022; pp.4611-4619.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.