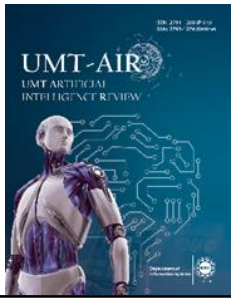
UMT Artificial Intelligence Review (UMT-AIR)
Volume 5 Issue 2, Fall 2025
ISSN$_{(P)}$: 2791-1276, ISSN$_{(E)}$: 2791-1268
Homepage: https://journals.umt.edu.pk/index.php/UMT-AIR



Title:	**Towards Robust and Explainable Early Rumor Detection: A Noise-Aware Graph Learning Framework**
Author (s):	Abu Bakar Shabbir and Usama Husnain
Affiliation (s):	University of Management and Technology, Lahore, Pakistan
DOI:	https://doi.org/10.32350.umt-air.52.01
History:	Received: July 21, 2025, Revised: October 13, 2025, Accepted: November 15, 2025, Published: December 03, 2025
Citation:	A. B. Shabbir and U. Husnain, "Towards robust and explainable early rumor detection: A noise-aware graph learning framework," *UMT Artif. Intell. Rev.*, vol. 5, no. 2, pp. 1–22, Dec. 2025, doi: https://doi.org/10.32350.umt-air.52.01.
Copyright:	© The Authors
Licensing:	This article is open access and is distributed under the terms of Creative Commons Attribution 4.0 International License
Conflict of Interest:	Author(s) declared no conflict of interest

A publication of
Department of Information System, Dr. Hasan Murad School of Management
University of Management and Technology, Lahore, Pakistan

Towards Robust and Explainable Early Rumor Detection: A Noise-Aware Graph Learning Framework

Abu Bakar Shabbir* and Usama Husnain

School of Systems and Technology, University of Management and Technology, Lahore, Pakistan

ABSTRACT Social media platforms have changed the communication process in such a radical way that now it takes just a few moments for unverified or fake news to reach millions of users. While in the past couple of years various methods have been developed to fight against the spread of false rumors, there are still many challenges to face. These challenges include dealing with noisy data, fast detection, and offering actionable insights for giving quick solutions. The current study aimed to put forward a conceptual noise-aware graph-based framework for the early detection of false rumors on social media. The framework melds together robustness, propagation-aware graph representations, and lightweight early-warning mechanisms in a single design. In order to depict the propagation patterns, the framework makes use of Graph Neural Networks (GNNs) for graph-based propagation modeling and also supports an uncertainty estimation module to handle data biases and incompleteness. One of the main features of the framework is the pinpointing of the most influential posts, users, and propagation paths that have a direct effect on the detection outcomes. While this work is not centered around experimental results, it lays out in detail the real-world validation plan using benchmark datasets, such as Twitter15/16, Weibo, and FakeNewsNet. The integration of noise-awareness, early detection, and interpretability in a single framework is the first step in the direction of robust and explainable rumor detection, thus opening new avenues for future theoretical and experimental work in this field.

INDEX TERMS early detection, explainability, Graph Neural Networks (GNNs), robustness, rumor detection, uncertainty

I. INTRODUCTION

The extensive use of social networking sites has significantly altered the way the world handles the flow of information, which is essentially produced, shared, and consumed. Though the social networking sites invite quick communication, these sites also facilitate the dissemination of unconfirmed and deceptive pieces of information. Falsehoods and fabrications on Twitter, Weibo, and Reddit can spread like wildfire within a few seconds, thus causing social panic, defamation, political polarization, and wrong public responses. The size, speed, and multiple modes of modern social media combined with the bot activity, cross-platform diffusion, and incomplete user metadata make manual fact-checking impossible and show the necessity of automated detection systems that may work in such noisy and rapidly-changing environments [1], [2].

Initially, rumor detection research relied on feature-based Machine Learning (ML) approaches (SVMs, decision trees, and Naïve Bayes) and has now evolved to Deep Learning (DL) models that can understand contextual and temporal signals [1], [3]. In

*Corresponding Author: abubakarshabbir64@gmail.com

the latest development, graph-based techniques, particularly Graph Neural Networks (GNNs) and graph-transformer architectures, have demonstrated a high level of performance as they represent the propagation structures, user interactions, and relational dependencies. Models, such as Bi-GCN and geometric DL methods are instrumental in both getting the right answer and in early detection by taking advantage of propagation-aware representations [4]–[6]. Similarly, evidence-aware GNNs and propagation-structure-aware transformers can be used for a higher level of robustness and interpretability [7], [8].

However, these improvements have left three important issues unanswered. Firstly, noise and uncertainty exist even in the most genuine and authentic social media data: propagation graphs are rarely complete, labels may be false, and user metadata might be scant or even tampered with. The robustness techniques currently in use have consequences that are only slightly better and are not specifically designed for the early detection task [6], [9]. Secondly, early detection is still a difficult task due to the fact that the initial propagation signals are very weak, incomplete, and extremely dynamic, thus making it necessary for models to derive conclusions from partial and evolving graph snapshots [5], [10]. Thirdly, explainability serves as a prerequisite for the introduction of the real world since platform moderators as well as policy-makers require understandable explanations of automated decisions. However, the majority of high-performing models still remain black boxes [1], [7], [8].

In an attempt to solve these problems, the current study put forward a consolidated noise-aware, graph-based framework for the early identification of rumors. The system combines: (i) noise modeling for uncertain labels, missing or spurious nodes/edges, and adversarial propagation patterns; (ii) a bi-directional propagation graph that integrates text, metadata, and sentiment-based node and edge features; and (iii) an explainable decision layer that delivers early-warning scores coupled with the interpretable traces of the most influential posts, users, and propagation paths [4], [11], [12].

Furthermore, this framework simultaneously tackles issues of noise-awareness, early detection, and explainability from a single conceptual pipeline. Unlike prior works that are primarily concerned with benchmark accuracy, the robustness of the proposed design in noisy conditions and the interpretability of the results for practical decision-making are its main features. The framework is based on robustness taxonomies [9], propagation-based modeling [4], [5], and multimodal evidence integration [7], [13] and at the same time, it proposes a well-structured, evaluation-ready approach for deployment in the real world.

Other parts of this study are structured as follows. Section 2 reviews the various rumor detection methods: traditional, DL, graph-based, multimodal, and noise-aware. Section 3 presents the early detection problem, noise models, and evaluation metrics. Section 4 elaborates on the features, graph building, detection modules, and explainability mechanisms of the proposed framework. Section 5 is a visual presentation of the case study of rumor propagation under the noise. Sections 6 and 7 outline the study limitations and experimental directions as well as deployment considerations. In sum, the framework is designed to provide robust, early, and interpretable rumor detection capabilities in contemporary

social networks.

II. LITERATURE REVIEW

The growing reach of social media platforms has changed the way information is shared. Thus, both verified and unverified contents have the same chance to be circulated to a large audience within seconds. Rumors are unverified or intentionally deceptive pieces of information that, by influencing the opinions and the decisions of the public, affect the social, political, and health sectors. As a result, the identification of rumors has become a vital target of the research, which was addressed at different levels and through various approaches. These include traditional ML, DL, graph-based models, multimodal fusion, noise-aware methods, and sentiment-informed approaches. In spite of advancements, there are still some major issues particularly concerning noisy propagation, early detection, and explainability. This part is dedicated to the review of the main methodological categories and their shortcomings.

A. TRADITIONAL MACHINE LEARNING (ML) APPROACHES

Traditional models were at the core of the first rumor detection systems that revolved around ML techniques. Support Vector Machines (SVMs), Decision Trees (DTs), and Naïve Bayes are examples of such systems. These methods focused on the manual creation of features from the text, user metadata, and the way the information was shared. In their work, Yang et al. [3] pointed out that features related to users were the most potent for differentiating rumors on Weibo. Additionally, Pathak et al. [1] conveyed that they achieved a performance level (up to 86%) on the datasets, such as KWON, MediaEval, and PHEME which was quite good. However, these conventional models, although they are interpretable and have low computational costs, are weak in capturing temporal dependencies, complex diffusion patterns, and multimodal signals which, in turn, restricts their capacity to be used in the early and stable detection of rumors.

B. DEEP LEARNING (DL) APPROACHES

DL techniques overcome the restrictions of manually crafted features by capturing linguistic, contextual, and temporal patterns directly from the data. RNNs, CNNs, and hybrid architectures have demonstrated better performance in rumor detection experiments [5], [10]. The Propagation Path Classification (PPC) model of Liu and Wu uses GRUs and CNNs to represent the spreading of a rumor as multivariate time series, thus it can detect very fast on Twitter15 within a few minutes. Monti et al. [5] also went beyond this by using geometric DL and the network structure to get a very high ROC-AUC score. Nevertheless, DL models that are based only on content or time-series data are still affected by noisy propagation, are not very robust to missing interactions, and usually have a small degree of interpretability.

C. GRAPH NEURAL NETWORK (GNN) APPROACHES

GNNs have been recognized as the most effective technique for detecting rumors, by appropriately representing the relational structures and diffusion dynamics. Song et al. [4] introduced the Bi-GCN model that combines the top-down and bottom-up propagation signals and thus, made a significant breakthrough in the performance on Weibo and Twitter15/16. Subsequently, more sophisticated models, such as the Propagation Structure-Aware Graph Transformer (PSGT) [8], help get rid

of the noise and facilitate interpretability by means of a self-attention mechanism, which is directed by the Information Bottleneck principle. The GET framework [7] is able to catch the intricate interactions between claim and evidence, thereby, achieving higher performance on the datasets, such as Snopes and PolitiFact. Although GNN-based approaches can achieve early detection and high performance, they typically demand stringent graph construction, heavyweight computational costs, and may still be sensitive to noisy and incomplete propagation graphs.

D. MULTIMODAL AND PROPAGATION-BASED METHODS

Multimodal and propagation-based approaches, as opposed to purely text-based models, aim to overcome the limitations of text-only models by incorporating various signals, such as textual content, images, metadata, and diffusion patterns. Zhang et al. [13] presented MKEMN, which integrates multimodal features with knowledge graphs to go beyond the semantic understanding. Shu et al. [11] introduced Hierarchical Propagation Networks (HPFN), which focus on macro- and micro-level diffusion aspects, and temporal features being very discriminative. While these techniques enhance generalization and allow for capturing a richer context, they are very demanding in terms of computational power, large training datasets are needed, and they mostly get worse when there is noise or incomplete propagation.

E. NOISE-AWARE AND UNCERTAINTY HANDLING IN RUMOR DETECTION

Noise-aware modeling is very important since social media data from the real world is noisy by nature and has posts that are mislabeled, edges that are missing, and propagation trees that are incomplete. Xu et. al. [9] examined various robustness mechanisms that include adversarial training, anomaly detection, preprocessing filters, and certifiable robustness for graph-based models. Recent architectures, such as PSGT [8], explicitly model noise in propagation patterns, which improves resilience to unreliable interactions. Uncertainty-aware methods (for instance, probabilistic embeddings or Bayesian GNNs) that are used for early detection of false positives can also be instrumental in alleviating the situation. This is because they provide a way to quantify the confidence level when the observations are incomplete.

F. SENTIMENT-AWARE APPROACHES

Sentiment signals can be very helpful to identify rumors since fake news frequently has an overly emotional tone. Alonso et al. [12] brought evidence that the sentiment polarity and intensity can be the factors that most differentiate true and false posts. Wang et al. [14] found that changes in sentiments had a strong impact on how people reacted to the COVID-19 rumors, thus pointing to the importance of the emotional and behavioral context. Even though sentiment-aware models make the system more understandable and can be used for very early detection, these models are not capable of handling the propagation structure on their own, so they need to be combined with graph-based features.

G. LIMITATIONS AND RESEARCH GAPS

While there has been a significant progress in ML, DL, GNN-based, multimodal, and sentiment-driven techniques, large gaps still remain. Traditional and first-

generation DL models are very dependent on feature engineering and are not quite robust to noisy propagation. GNNs and graph transformers provide better structural modeling but require complete, high-quality graphs and are usually not easily interpretable. Multimodal and sentiment-aware methods extract more features from the data but have a computational overhead and are sensitive to missing modalities. Most importantly, in most cases, the existing methods rarely integrate noise-awareness, early detection, and explainability in a single framework. This discrepancy is the reason for the noise-aware graph-based framework proposal which combines robustness, timeliness, and interpretability to detect rumors in a noisy social media environment effectively. Table I gives an overview of the reviewed approaches and provides a summary of how each approach contributes to the understanding of the challenges in noise-aware early rumor detection.

TABLE I
SIMPLIFIED COMPARATIVE SUMMARY OF RUMOR AND FAKE NEWS DETECTION STUDIES

Author and Year	Objective/Problem	Method/Approach	Key Results
Pathak et al. [1]	Survey rumor detection datasets	ML (SVM, DT, NB), DL (RNN, CNN), Hybrid	ML ≈ 86%, DL > 90%
Phan et al. [6]	Analyze GNN-based fake news detection	Survey of 27 GNN models (GCN, AGNN, GAE)	Up to 20% better than ML
J. Xu et al. [9]	Address GNN vulnerability to noise and adversaries	Robustness survey: anomaly detection, adversarial training, defenses	Robustness taxonomy
Monti et al. [5]	Propagation-based fake news detection	Geometric Deep Learning (Graph CNNs)	AUC ≈ 92.7%
Song et al. [4]	Rumor detection on noisy graphs	Bi-GCN + DropEdge	Acc: 96% / 88%
Shu et al. [11]	Hierarchical propagation modeling	HPFN (macro–micro propagation networks)	F1 > 0.84
Liu & Wu. [10]	Early fake news detection	PPC (GRU + CNN)	Early acc ≈ 92%
Yang et al. [3]	Weibo rumor detection	SVM + crafted features	+6.3% improvement
Zubiaga et al. [2]	Rumor detection + resolution survey	ML/NLP 4-stage rumor framework	Structured rumor pipeline
Zhang et al. [13]	Multimodal fake news detection	MKN + EMN (text + image + KG)	Acc 86% / 81%
W. Xu et al. [7]	Evidence-aware detection	GET (graph semantic mining)	SOTA F1
Zhu et al. [8]	Robust + interpretable detection	PSGT (Graph Transformer + Info Bottleneck)	F1 ≈ 0.98
Alonso et al. [12]	Sentiment role in fake news	Lexicon/ML/DL sentiment models	Acc ≈ 0.94
Khazane et al. [15]	Adversarial ML in IoT	Survey taxonomy of attacks/defenses	Structured taxonomy
Wang et al. [14]	Rumor–sentiment dynamics	Topic modeling + SA + ML	F1 = 0.91

H. PROBLEM FORMULATION

The rapid expansion of platforms, such as Twitter, Weibo, and Facebook has transformed information creation and consumption, however, it has also accelerated the spread of misinformation and rumors [1], [4], [6]. Unverified rumors or intentionally misleading claims may trigger social, political, and economic harm. Automatic rumor detection is challenging since social media data exhibits noisy content, incomplete propagation, heterogeneous user behavior, and multimodal signals.

A social media environment can be modeled as a collection of posts $\mathcal{P} = \{p_1, \dots, p_n\}$ and users $\mathcal{U} = \{u_1, \dots, u_m\}$. Each post p_i contains text and possibly images, videos, metadata (timestamps, hashtags), and engagement signals (likes, comments, retweets). The spread of post p_i is represented as a propagation graph $G_i = (V_i, E_i)$, where nodes represent posts or users involved in dissemination and edges represent interactions (retweets, replies, mentions, shares). Each node $v \in V_i$ has a feature vector $\mathbf{x}_v$ describing user attributes (followers, account age, verification), linguistic embeddings, and sentiment cues. Edges $e \in E_i$ encode interaction type and temporal delays.

I. FORMAL DEFINITION OF RUMOR DETECTION

Given the propagation graph G_i of a post p_i, rumor detection is formulated as assigning a binary label:

$$y_i \in \{0,1\}, \quad y_i = 1 \text{ (rumor)}, \quad y_i = 0 \text{ (non-rumor)}.$$

The detection function is

$$f: G_i \rightarrow y_i.$$

However, real-world graphs are often noisy or partially observed. Let $\tilde{G}_i$ denote such a noisy propagation graph (with missing nodes/edges or corrupted features). The problem becomes a noise-aware classification task:

$$f: (G_i, \tilde{G}_i) \rightarrow y_i.$$

J. EARLY DETECTION CONSTRAINTS

Early rumor detection requires predicting the label using partial propagation snapshots G_i^t, where t represents the elapsed time since initial posting. Challenges include:

- **Incomplete Propagation:** Only a small portion of G_i is revealed early, limiting structural cues [5], [10].
- **Sparse Interactions:** Few retweets or replies make early classification difficult.
- **Real-time Constraints:** Effective systems must operate with low latency to issue timely warnings.

The early detection objective is:

$$f: G_i^{te} \rightarrow y_i, \qquad t_e \ll T,$$

where T is the full observation period.

Early detection performance is often measured using Early Detection Score (EDS), which rewards both correctness and timeliness.

K. NOISE MODELING IN SOCIAL MEDIA GRAPHS

Real social media propagation is inherently noisy. The following noise types commonly appear:

1) LABEL NOISE

Fact-checking delays or subjective annotation cause incorrect labels. The observed label $\hat{y}_i$ differs from the true label y_i with probability:

$P(\hat{y}_i \neq y_i) = \epsilon$.

2) NODE NOISE

User features may be incomplete or manipulated (e.g., bots). Corrupted node features are modeled as:

$\tilde{\mathbf{x}}_v = \mathbf{x}_v + \delta_v$.

3) EDGE NOISE

Missing or spurious interactions distort propagation:

$E_i \backslash \tilde{E}_i$ (missing edges), $\tilde{E}_i \backslash E_i$ (spurious edges).

4) INCOMPLETE PROPAGATION

Early graph snapshots G_i^t reveal only a subset of nodes and interactions, producing structural uncertainty.

As illustrated in Figure 1, social media propagation is affected by multiple noise sources that reduce stability and early detection accuracy. Noise-aware models incorporate these uncertainties via probabilistic embeddings, uncertainty propagation, or robust graph convolutions [8], [9]. Table II summarizes the main types of noise in social media graphs, their definitions, and their impact on rumor detection.

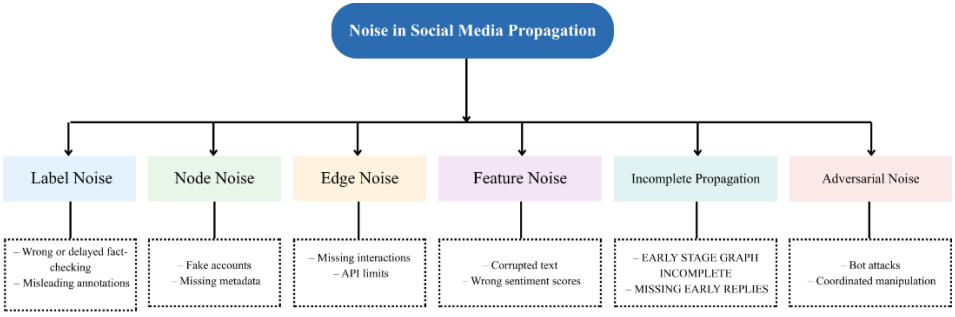


FIGURE 1. Taxonomy of noise sources in social media propagation, including label noise, node noise, edge noise, feature noise, incomplete propagation, and adversarial noise.

TABLE II
TYPES OF NOISE IN SOCIAL MEDIA GRAPHS AND THEIR IMPACT ON RUMOR DETECTION

Noise Type	Definition	Impact on Rumor Detection	Real-World Example
Label Noise	Incorrect or inconsistent labels due to delayed fact-checking or subjective annotation.	Misleads learning, increases false positives/negatives, reduces model stability, especially in early detection.	A tweet initially labeled "true" but later verified as false.
Node Noise	Incomplete or unreliable user/post features from bots, missing metadata, or manipulated profiles.	Weakens node embeddings and propagation semantics, distorting GNN feature aggregation.	Bot accounts with unrealistic follower ratios or missing profile data.

Noise Type	Definition	Impact on Rumor Detection	Real-World Example
Edge Noise	Missing or spurious interactions (retweets, replies, mentions) due to API limits or manipulation.	Distorts graph structure, weakens connectivity, and disrupts diffusion modeling.	Missing retweet edges during high-traffic events due to API rate limits.
Feature Noise	Corrupted or incomplete text, sentiment, temporal, or metadata features.	Causes unstable feature extraction and incorrect learning of semantic/temporal patterns.	Sarcastic or multimedia-dependent posts producing inaccurate sentiment scores.
Incomplete Propagation	Partial observation of the propagation graph during early stages.	Limits structural cues, increases uncertainty, and makes early-stage prediction harder.	Only a few replies/retweets visible in the initial minutes of an event.
Adversarial Noise	Coordinated manipulation by bots or trolls injecting misleading content or propagation signals.	Confuses the model, amplifies false patterns, and reduces robustness.	Botnets rapidly boosting a politically motivated rumor.

L. EVALUATION METRICS

Rumor detection performance is typically assessed using standard metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

1) EARLY DETECTION SCORE (EDS)

Let T_i denote the full observation horizon for event i and t_i the earliest correct prediction time with confidence above threshold τ. Define:

$$\text{EDS}_i = \begin{cases} 1 - \frac{t_i}{T_i}, & t_i \le T_i, \\ 0, & \text{otherwise.} \end{cases}$$

The aggregated score is:

$$\text{EDS} = \frac{1}{N} \sum_{i=1}^{N} \text{EDS}_i.$$

2) ROBUSTNESS METRIC (RM)

Given baseline metric M_0 (e.g., F1) and noisy-data metric $M(\eta)$ under noise level η:

$$\text{RM}(\eta) = \frac{M(\eta)}{M_0}.$$

Values close to 1 indicate high robustness. Table III summarizes the metrics used to evaluate the proposed framework, including standard classification, early detection, and robustness measures.

M. CHALLENGES IN NOISE-AWARE EARLY RUMOR DETECTION

Noise-aware early rumor detectors face a number of significant problems that have to be solved before they can be efficiently used. To begin with, the propagation graph in the very early stages is usually incomplete, so there are missing nodes and

edges that hide the real structural patterns necessary for strong reasoning. Besides that, rumor spreading is a variable process by nature, so the models have to be updated with temporal changes instead of using the static structure assumptions. Moreover, the framework should be able to work with different and multimodal features. In this case, some node or edge attributes may be missing, that is to say unreliable or partially observable. The most difficult problem, however, is due to adversarial manipulation when bots, coordinated campaigns, or malicious users deliberately and consistently introduce misleading noise that is not random into the propagation process. The last thing to mention is that the recognition tool has to be as clear as possible and provide supported explanations about those nodes, edges, or propagation ways that influenced the decision and, thus, facilitate human examination, accountability, and trust in automated decisions.

TABLE III
EVALUATION METRICS FOR NOISE-AWARE EARLY RUMOR DETECTION

Metric	Formula	Purpose	What it Measures
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness	Proportion of correct predictions
Precision	$\frac{TP}{TP+FP}$	Correctness among positive predictions	Fraction of predicted rumors that are true
Recall	$\frac{TP}{TP+FN}$	Completeness of detection	Fraction of actual rumors detected
F1-score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall	Balances precision and recall
Early Detection Score (EDS)	$EDS = \frac{1}{N}\sum_{i=1}^{N} E\,DS_i$; 1 if correct, 0 otherwise	Temporal responsiveness	Rewards early correct predictions
Robustness Metric (RM)	$RM(\eta) = \frac{M(\eta)}{M_0}$	Performance under noise	Drop in performance under noisy conditions

N. PROBLEM SUMMARY

The noise-aware early rumor detection task is formulated as learning:

$f_\theta: \tilde{G}_i^t \rightarrow \hat{y}_i$,

where $\tilde{G}_i^t$ is a noisy, partial propagation graph at time t, θ are model parameters, and $\hat{y}_i$ is the predicted label. An effective model must:

- handle noise from labels, nodes, and edges;
- make precise predictions at the very beginning (t≪T);
- give understandable explanations of the most influential nodes and the propagation paths;
- become very large, dynamic social

networks.

It necessitates the use of graph representation learning, uncertainty modeling, and temporal propagation analysis to solve this problem. The proposed noise-aware framework that merges these elements is thus an instantiation of a comprehensive system capable of robust, early, and interpretable rumor detection in the presence of noise.

III. PROPOSED FRAMEWORK

The study presented a single unified noise-aware graph-based framework for early rumor detection which is capable of solving the three problems that most of the time arise in the area of misinformation detection: (i) noisy or incomplete social media data, (ii) early-stage detection with partial propagation, (iii) interpretation for trustworthy decision-making. This framework combines GNNs, uncertainty modeling, and temporal propagation analysis into one pipeline that is specifically designed and is efficient in real-world noisy scenarios.

Table IV gives a detailed breakdown of the components of the proposed noise-aware graph-based framework with columns for inputs, outputs, purposes, and techniques of each layer.

TABLE IV
COMPONENTS OF THE PROPOSED NOISE-AWARE GRAPH-BASED FRAMEWORK FOR EARLY RUMOR DETECTION

Layer	Input	Output	Purpose	Techniques
Input Layer	Raw posts, metadata, propagation graphs, optional modalities	Feature-enriched nodes/edges	Collects and structures multi-source data	Text embeddings, metadata extraction, sentiment scoring
Preprocessing and Noise Handling Layer	Raw node and edge features	Cleaned and normalized features	Removes noise, handles missing data	Text cleaning, node normalization, imputation, data augmentation
Graph Representation Layer	Node/edge features	Structured graphs for GNN input	Converts data into structured graph format	Bi-directional graph construction, temporal encoding
GNN Architecture	Graphs with features	Node embeddings	Learns semantic and structural representations	GCN/GAT/Graph Transformer with reliability-weighted aggregation
Noise-Aware Rumor Detection Layer	Node embeddings	Rumor probability scores	Detects rumors considering uncertainty	Bayesian GNNs, MC dropout, robust loss

Layer	Input	Output	Purpose	Techniques
Decision & Early Alert Layer	Rumor probability scores	Early alerts and flagged posts	Generates explainable early warnings	Adaptive thresholds, explainability visualization
Conceptual Evaluation Layer	Model outputs	Evaluation metrics	Quantifies performance	Accuracy, F1, EDS, Robustness, Explainability metrics

A. INPUT LAYER

The framework ingests heterogeneous social media signals that collectively support robust graph construction and representation learning:

- **Text Content:** Raw posts (tweets, replies, retweets) forming the primary semantic information.
- **Metadata:** These are the features of the users (how old an account is, how many followers does the user have, whether the user is verified), timestamps, and platform-specific indicators.
- **Propagation Structure:** The retweet, reply, and mention interactions that capture the flow of the information.
- **Optional Modalities:** Pictures, videos, and sentiment scores if available.

Such inputs allow for a complete modeling of the content and the context which are the core of the early rumor detection.

B. PREPROCESSING AND NOISE HANDLING LAYER

Considering the disruptive aspects of the real-world social media data, specific noise-handling mechanisms have been implemented:

- **Text Normalization:** Removal of URLs, hashtags, emojis, and non-linguistic symbols; tokenization and lowercasing for downstream embedding.
- **Node Feature Normalization:** Conversion of non-categorical attributes (e.g., follower count) to a different scale to avoid the bias of the magnitude.
- **Noise-Aware Propagation Modeling:** Derivation of node/edge trustworthiness from metadata completeness, activity, and propagation consistency.
- **Imputation and Augmentation:** Exploiting the graph to impute missing node attributes and artificially generating low-activity propagation sequences for sparsity reduction.

By explicitly modeling uncertainty, the system is less susceptible to incomplete or corrupted propagation.

C. GRAPH REPRESENTATION LAYER

Different raw multimodal data are converted into structured propagation graphs:

- **Node Features:** Text embeddings, user statistics, sentiment scores, and if necessary, visual descriptors.
- **Edge Features:** Interaction type, frequency, and temporal recency.
- **Bi-directional Graph Construction:**

The graph not only shows top-down diffusion (source to retweets) but also bottom-up engagement (user feedback to source) hence, it traces both the influence flow and the interaction context.

- **Temporal Encoding:** The time intervals are the edges through which the model can get the propagation speed that is very much needed for early detection.

The graph structure here is a kind of storage for relational dependencies that differentiate rumors from non-rumors.

D. GNN ARCHITECTURE

Let $\mathbf{h}_v^{(l)}$ denote the embedding of node v at layer l. We introduce reliability-weighted message passing to account for noisy edges:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(W^{(l)} \sum_{u \in \mathcal{N}(v)} \tilde{r}_{uv} \frac{\mathbf{h}_u^{(l)}}{\sqrt{\tilde{d}_u \tilde{d}_v}} \right),$$

where $\tilde{r}_{uv} = r_{uv} / \sum_{u' \in \mathcal{N}(v)} r_{u'v}$ is the normalized reliability, $\tilde{d}_v = \sum_{u \in \mathcal{N}(v)} r_{uv}$, and σ is an activation function.

Edge reliability is parameterized as:

$$r_{uv} = \text{sigmoid}(w_1 s_u + w_2 \,\text{freq}_{uv} + w_3 \,\text{recency}_{uv}),$$

where s_u is node reliability, freq_{uv} is interaction frequency, and recency_{uv} is a temporal decay factor. Learnable parameters w_1, w_2, w_3 adapt reliability estimation to data conditions.

E. NOISE-AWARE RUMOR DETECTION LAYER

The rumor detection layer integrates GNN outputs with uncertainty-aware mechanisms:

- **GNN or Graph Transformer Backbone:** Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), or Graph Transformers depending on computational requirements and diffusion complexity.
- **Uncertainty Modeling:** Bayesian GNNs or MC dropout quantify prediction uncertainty in the presence of noisy or missing data.
- **Dynamic Early Scoring:** Node-level rumor likelihood is updated as new posts arrive; adaptive thresholds support early detection.

F. UNCERTAINTY ESTIMATION AND ROBUST LOSS

Predictive distributions were obtained through S stochastic forward passes:

$$\bar{p}_c = \frac{1}{S} \sum_{s=1}^{S} p_c^{(s)}, \qquad H(\bar{p}) = -\sum_c \bar{p}_c \log \bar{p}_c.$$

The early alert rule is:

flag if $\bar{p}_{\text{rumor}} > \tau_p$ and $H(\bar{p}) < \tau_H$.

To mitigate label noise, we adopt Generalized Cross-Entropy (GCE):

$$\mathcal{L}_{\text{GCE}}(p, y) = \frac{1 - p_y^q}{q}, \qquad q \in (0,1].$$

The full objective is:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \omega_i \; \mathcal{L}_{\text{GCE}}(p_i, y_i) + \lambda \, \mathcal{L}_{\text{consistency}},$$

where ω_i encodes event reliability and $\mathcal{L}_{\text{consistency}}$ enforces stability under graph augmentations.

G. DECISION AND EARLY ALERT LAYER

Once the prediction phase is over, the decision layer incorporates an adaptive thresholding technique that relies on uncertainty estimation to decrease the number of false alarms, in particular, those that arise in the first stages of the propagation. Thereafter, the system generates understandable results with the assistance of saliency-based or attention-driven explanation methods, which give the nodes, edges, and propagation paths that influenced the model's decision the most. In the end, the framework, in essence, if the computed rumor probability is above the threshold, turns on the early warning system to notify the human analysts for a quick intervention. To put it briefly, this layer delivers a tightly coordinated effort of accuracy, interpretability, and operational transparency, which, for instance, in the case of safety-critical real-world applications, is absolutely necessary.

H. CONCEPTUAL EVALUATION LAYER

The proposed framework undergoes testing with a wide range of metrics that, in combination, reflect the detection quality, timing, resilience, and interpretability. Besides the classification metrics of accuracy, precision, recall, and F1-score that describe the general predictive capability, the set of metrics is also advanced by early detection metrics. Hence, time-to-detection, EDS, and early-stage recall are the terms used to express the sensitivity at the very first propagation stages. The robustness side of the coin is a performance drop due to noise that has been deliberately injected, for instance, missing edges or corrupted labels, whereas explainability is gauged by fidelity and sparsity metrics for the most influential components. These criteria form a basis for a strict comparison with the state-of-the-art methods, such as Bi-GCN [4], PSGT [8], MKEMN [13], and HPFN [11].

The proposed architecture is depicted in Figure 2, which shows the major components of the noise-aware early rumor detection pipeline.

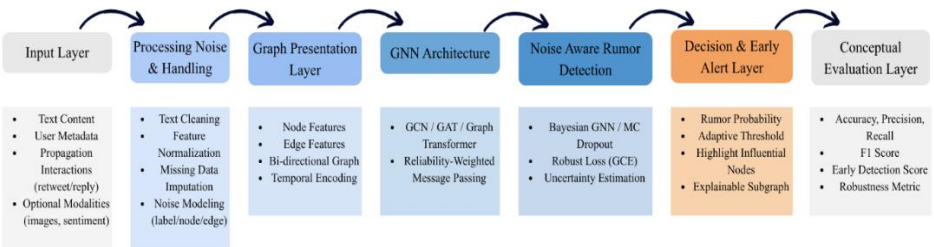


FIGURE 2. The overall noise-aware graph-based framework for early rumor detection, showing input processing, noise handling, graph construction, GNN reasoning, uncertainty modeling, and early alert generation.

I. INTEGRATION AND WORKFLOW SUMMARY

The detailed workflow of the framework moves sequentially from data ingestion, through preprocessing and noise modeling operations, to the construction of temporal

and bi-directional graph structures and finally noise-aware inference. At the beginning, data from various social media streams is gathered. The data then goes through the preprocessing and noise modeling stages, which aim to clean, normalize, and estimate the reliability of the input data. After that, temporal and bi-directional graph structures are built, and noise-aware inference is carried out with the help of GNN or Transformer-based architectures. The rumor likelihood scores are changed dynamically over time, thus giving the possibility of early detection. Moreover, the system offers explainable decision support by finding the most influential nodes and the propagation pathways and ends with a thorough evaluation in terms of accuracy, timeliness, robustness, and transparency. The approach, in this respect, is conceptually similar to the commonly used misinformation datasets, such as Twitter15/16, Weibo, and FakeNewsNet.

J. CONCEPTUAL ADVANTAGES

The proposed architecture comes with numerous benefits. To start with, noise-aware is achieved through reliability-weighted message passing and uncertainty modeling, which enhances resistance to missing or corrupted data. Secondly, temporal encoding and dynamic scoring allow for the detection of very early stages, thus, intervention can take place before rumors become widely known. Thirdly, the use of graph-based learning allows the model to deeply understand relational and temporal dependencies that are infeasible for purely content-based methods. Moreover, the system keeps interpretability by identifying the most influential users, posts, and edges, thus, it is a good fit for content moderation which is of a non-transparent nature. Lastly, the layout is also capable of handling multimodal data sources and can be referred to as different social media platforms, thus, it is able to support different operational contexts.

IV. ILLUSTRATIVE EXAMPLE/CASE STUDY

The study put forth a noise-aware graph-based framework that can be implemented in reality by presenting a made-up scenario that is still quite representative of the way rumors typically spread on social media. The purpose of this case study was to display how the framework operates as it deals with noisy inputs, builds propagation structures, accounts for uncertainty, and produces interpretable early warnings.

A. SCENARIO DESCRIPTION

Imagine a situation where an unsubstantiated report goes viral on a widely used microblogging service, saying that the deterioration of the local health condition is the government's fault, which has been caused by negligence. The very first post is made by a moderately influential user (User A), and as a result, the message gets propagated quickly as users with different levels of credibility retweet, reply, and comment. In a short period of time, a live propagation graph comes into existence but the text also has a number of noise elements, such as incomplete metadata, missing sentiment information, inconsistent linguistic cues, and partially correct or wrong retweet chains. The main issue here is to pinpoint the rumor in those early, noisy propagation stages while also making sure that the model's decision is interpretable.

B. GRAPH CONSTRUCTION AND PREPROCESSING

The framework restructures the initial exchanges into a graph of structured propagation, where every node stands for a user-post pair that includes textual

embeddings, user metadata, and sentiment features. The relationships here are retweet and reply ones, and a bi-directional configuration reflects both the top-down diffusion from the source post and the bottom-up engagement patterns from responding users. Any missing or corrupted features, for instance, missing sentiment scores or incomplete user metadata, are restored with the help of statistical and graph-based methods, whereas noisy edges related to bot-like activities or unusually fast replies are given lower weights by reliability scores calculated from the behavioral history of the respective accounts. Textual data undergoes standardization and embedding with the help of language models which have already been trained, thus linguistic noise is minimized, and semantic meaning is retained. A reduced version of the first propagation network is presented in Figure 3, where the source post, the influential early users, and the noisy bot clusters are emphasized.

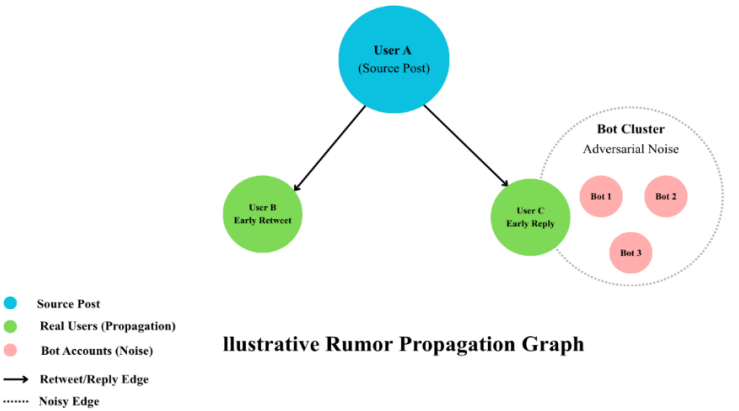


FIGURE 3. Example of an early-stage rumor propagation graph showing the source post, early amplifiers, and noisy bot-like interactions.

C. PROPAGATION ANALYSIS AND EARLY DETECTION

After building the graph, the noise-aware GNN considers the features of the nodes and edges while also keeping track of the uncertainty due to the unreliable information. A confidence score reflecting the trustworthiness of the features is assigned to each node. To illustrate, an account that was just created, and is posting inconsistent or emotionally charged content would get a lower confidence score, thus its influence on the final decision may be minimal. The model updates the estimates of rumor likelihood at each step of the propagation almost in real-time. Here, the combined rumor probability goes over the decision threshold shortly after ten interactions, thus giving the notification of an early alert well before the rumor gaining widespread visibility. It, therefore, shows that the model is capable of detecting emerging rumors even when the propagation is sparse and there is a lot of noise.

D. NOISE-AWARE MODELING

The framework is always ready to deal with

various kinds of noises during the process of inference. To cope with label noise, the system uses uncertainty-aware loss functions that lessen the effect of ambiguous or difficult training samples. The noise in the structure that comes from missing or spurious edges is solved by probabilistic edge weighting that changes the trust based on the interaction history and user reliability. The feature noise, mainly in textual, metadata, and sentiment attributes is controlled through attention mechanisms that select the most consistent and informative neighbors and at the same time lessen the dependence on the low-confidence nodes. All these methods together guarantee that the detection performance would be at a stable level even if the noise is very high.

E. EXPLAINABILITY AND INTERPRETABILITY

The distinctive characteristic of the framework is its capability to offer clear explanations for its predictions. Once the system detects the rumor, it looks for the nodes and edges that had the most influence and led to the conclusion. It recognizes the original post of User A as the source of the rumor while emphasizing the roles of Users B and C through their early and highly impactful interactions. The framework also points out that a small group of bot-like accounts quickly and repetitively replying helped the propagation be noisy. The subgraph visualization based on these results provides the analysts with a brief, clear, and to the point view of the rumor dissemination, the significant actors, and the noisy parts which affected the detection process.

F. STEPWISE NARRATIVE OF THE DETECTION PROCESS

At first, rumor detection entails extracting the main text and metadata from a post of the first user (User A). The prompt reactions from users (Users B, C, and D) along with the graph structure also update and thus, each new communication node recalculates the confidence and reliability scores for edges. The missing sentiment for a few users is provided via imputation, whereas very quick or suspicious retweets are given low trust levels. When the GNN disseminates data among the nodes, uncertainty-aware reasoning determines the rumor probability that eventually exceeds the early-warning threshold. After a warning signal is sent out, the system displays the most influential posts and the closest links so that moderators or researchers can easily confirm them.

G. ILLUSTRATIVE OUTCOMES

The case depicts that the architecture is able to pinpoint a rumor at a very early stage of its life cycle, usually, it is done prior to the depth of significant propagation. Thanks to the noise-aware design, the system has the ability to reject those interactions which mislead or are artificially-amplified, thus it comes up with consistent predictions even if there is misinformation, incomplete metadata, or platform irregularities. Besides that, interpretability mechanisms are always at hand for decision-makers in order to get the understanding of the basis of each alert which, in turn, helps to raise trust, accountability, and the proper way of the intervention. The use of bi-directional propagation modeling additionally gives the detailed insight of how the flow of information from the source to the audience and vice versa happens within the network. In general, the example, though made up, is very close to the real behavior of social media and it serves as a demonstration of the practical application of the proposed framework.

V. DISCUSSION

The noise-aware graph-based framework for early rumor detection put forward by the authors installs the concept of structural graph learning, noise modeling, uncertainty estimation, and interpretable decision outputs in order to solve the problem of misinformation that is currently unaddressed in the literature. The compositional design is inspired by the strengths of recently introduced graph-based methods (e.g., Bi-GCN, PSGT, GET) but is aimed at three tightly coupled targets, which are not usually achieved at the same time in previous works: 1) robustness to noisy and incomplete propagation 2) early-warning capability, and 3) actionable interpretability. The following discussion portrays the main features and the potential of the proposed method, balances the limitations and compromises, and sketches the directions of the further experimental and deployment works for the purpose of confirming and extending the study.

A. STRENGTHS AND CONTRIBUTIONS

The proposed framework offered several notable strengths, foremost among them its comprehensive treatment of multiple noise sources including label, node, and edge noise through reliability-weighted aggregation, probabilistic edge modeling, and uncertainty-aware loss functions. This integrated approach reduces the heavy dependence on fully accurate propagation graphs that constrains many prior models. By assigning confidence scores to nodes and edges and by applying imputation and augmentation during preprocessing, the framework better reflects the imperfect conditions typical of real-world social media ecosystems. One of the significant features of the system is that it implements probabilistic predictive distributions, such as Monte Carlo dropout. Furthermore, it uses decision rules involving both predictive mean and entropy to provide an adaptive early-flagging mechanism which can quite adequately make the timeliness and reliability trade-off. Thus, it solves the problem of a large number of false positives that are costly in the earliest stages of rumor spread. In addition to this, graph construction in both directions and temporal encoding not only assists in capturing top-down as well as the bottom-up influence relationships but also provides more comprehensive structural cues to the model than just those from content or uni-directional methods [4]. Moreover, a stable GNN or Graph Transformer architecture coupled with a strong stochastic loss function like Generalized Cross-Entropy and consistency regularization on perturbed graphs can significantly stabilize the model's performance in noisy and dynamic environments. As a matter of fact, the model is a step ahead and identifies the most influential nodes, edges, and subgraphs that cause its predictions [8], [12], thus enabling analysts and moderators to receive understandable insights rather than a black-box model.

B. EVALUATION AND METRIC SUITABILITY

The set of evaluation metrics put forward for the framework matches the latter's goals quite closely. Standard classifying metrics, such as accuracy, precision, recall, and F1-score serve to measure the overall predictive power. Whereas, EDS is a metric that temporally gauges responsiveness and it also heavily penalizes those instances of late detections that standard metrics overlook. The robustness metric RM(η), which refers to a model's capability at various noise levels is very helpful in systematically comparing different noise-handling approaches. Explainability

metrics that gauge the extent to which the identified nodes or subgraphs correspond to the ground-truth propagation drivers provide a feasible manner of assessing the practical usefulness of the model's interpretability. These metrics may be combined and shown to the experimental participants. For instance, F1-versus-EDS curves under various noise conditions can be used to demonstrate the trade-offs among accuracy, timeliness, and robustness and to reveal the multidimensional strengths of the proposed framework.

C. LIMITATIONS AND TRADE-OFFS

Though the framework has a number of advantages, it also reveals a handful of disadvantages and trade-offs, which should be acknowledged. The noise-aware components and uncertainty estimation that the framework commits to lead towards an increased computational complexity, especially since Monte Carlo dropout-like segments calling for multiple forward passes are inherently slow. On the other hand, reliability-weighted aggregation and detailed edge modeling increase per-edge computational costs as well, and this can become a non-trivial problem when working at the scale of social media. Furthermore, the method depends on edge-reliability heuristics or learnable scalars, and bad calibration or domain shift between platforms may cause performance to drop, which is a problem that has been illustrated by DDT [16], a domain-aligned rumor detection paper. Though imputation and synthetic augmentation may lessen the sparsity of data, they are at risk of introducing bias if the generated propagation patterns are not very close to the actual ones. Such a predicament was acknowledged in multimodal augmentation methods [13] previously. Ultimately, despite the fact that the framework yields interpretable subgraphs and attribution maps, the extent to which these outputs can be understood in the real world is mainly dependent on the visualization brand and the skill level of the analysts. Numerical explainability scores do not by any means facilitate operational clarity.

D. PRACTICAL CONSIDERATIONS FOR SCALABILITY AND DEPLOYMENT

Moving the framework from a conceptual to an operational stage would require dealing with scalability issues as well as constraints related to the platform. Large-scale propagation graphs would require mini-batch training coupled with neighborhood sampling for the training process, while incremental updating of node embeddings instead of full-graph recomputation could be a way to considerably lower inference latency during early detection. Model compression and early-exit mechanisms are examples of methods that may limit the computational cost by enabling simpler sub-models to work in low-uncertainty cases. Per-platform calibration is also needed for a robust deployment, for instance, tuning reliability estimators, temporal decay parameters, and user-behavior priors to adjust for differences between networks, such as Twitter and Weibo. Last but not least, moral and legal aspects call for data handling that preserves privacy, thus aggregated or anonymized features being preferred over raw user data whenever that is feasible.

E. EXPERIMENTAL VALIDATION ROADMAP

On the one hand, a well-thought-out experimental program is a must to confirm the claims of the proposed framework. It should comprise various evaluation dimensions. On the other hand, benchmark comparisons performed on Twitter15/16,

Weibo, and FakeNewsNet may serve as an initial test of performance and present the baseline when using F1-score, EDS, and ROC or robustness curves jointly. Followingthis, the framework's strength should be verified under systematically induced noise, for instance, label corruption, missing edges, and synthetic bot activity. Ablation experiments are absolutely necessary to gauge the effect of core components, such as reliability-weighting, uncertainty estimation, and the employment of a robust loss function. Efficiency profiling should be about detection latency, computation time, and resource usage during both training and inference in order to decide whether real-time deployment is feasible. Furthermore, the use of the human-in-the-loop for evaluation, whereby analysts are requested to review explanatory subgraphs, can provide insight into the practical utilization of the interpretability features. Cross-domain transfer experiments employing data from one platform to train a model and data from another platform to test it may be the final point of the line. In this way, the degree of generalizability and robustness to domain shift can be figured out. These tests, if combined, provide a thorough and transparent account of the framework's effectiveness, efficiency, and applicability in the real world.

F. EXTENSIONS AND FUTURE RESEARCH DIRECTIONS

The current study has brought about many promising directions for the future. The fusion of multiple data types is the natural next step to extend the propagation graph with visual features, shared URLs, and knowledge-graph signals to capture more in-depth contextual cues. Active learning strategies could be employed to annotate high-uncertainty events as a priority thus, label noise would decrease gradually. Learning approaches that are federated or privacy-preserving can be a solution to the problem of model adaptation across platforms without the need for sharing raw user data which is in line with the regulators' and ethicists' expectations. The next generation of systems may also feature inclusion of adversarial training or certifiable robustness techniques to fight against malicious coordinated manipulation and sophisticated bot networks thus, the framework's resilience to adversarial environments would be fortified.

G. ETHICAL AND OPERATIONAL IMPLICATIONS

The introduction of an early-warning system for misinformation is a big step that should be accompanied by ethical responsibilities. On the one hand, false positives may lead to the suppression of legitimate speech, on the other hand, false negatives may give a chance to the harmful content to be spread without any control. Uncertainty quantification and explainability are the framework's features that may help with the safe implementation of the system by enabling confidence-aware routing of predictions to human validators and by giving transparent justifications for downstream interventions. However, these features should be governed within a framework that provides, among other things, accountability mechanisms, transparency of model decision criteria, periodic bias audits, and drift monitoring. The latter may help ensure that the deployment of early rumor detection systems stays consistent with society's notions of fairness, accountability, and trust.

H. CONCLUSION

This study’s authors developed a comprehensive and principled framework

for early rumor detection in which these three capabilities were combined: (i) explicit modeling of noise, (ii) uncertainty quantification, (iii) the use of temporal graph representations, and (iv) the provision of interpretable outputs. The conceptual design, as such, looks very promising for early robust detection. However, it still needs to be confirmed that the system performs well on standard benchmarks (e.g., Twitter15/16, Weibo, FakeNewsNet) and under controlled noise injections, that it can be efficiently optimized for real-time inference, and that governance controls (thresholding policies, human-in-the-loop triage, auditing) can be integrated for dealing with false positives and/or misuse. As a result of these updates, the framework would not only be more user-friendly in real-life scenarios but also, it could notably facilitate the work of moderators and analysts [8].

Author Contribution

Abu Bakar Shabbir: conceptualization, methodology, software, investigation, formal analysis, validation, visualization, writing – original draft. **Usama Husnain:** conceptualization, methodology, resources, data curation, writing – review & editing.

Conflict of Interest

The authors of the manuscript have no financial or non-financial conflict of interest in the subject matter or materials discussed in this manuscript.

Data Availability Statement

Data supporting the findings of this study will be made available by the corresponding author upon request.

Funding Details

No funding has been received for this research.

Generative AI Disclosure Statement

The authors did not used any type of generative artificial intelligence software for this research.

REFERENCES

[1] A. R. Pathak, A. Mahajan, K. Singh, A. Patil, and A. Nair, "Analysis of techniques for rumor detection in social media," *Procedia Comput. Sci.*, vol. 167, pp. 2286–2296, Mar. 2020, doi: https://doi.org/10.1016/j.procs.2020.03.281.

[2] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, June 2018, doi: https://doi.org/10.1145/3161603.

[3] F. Yang, X. Yu, Y. Liu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2012, pp. 1243–1251, doi: https://doi.org/10.1145/2350190.2350203.

[4] S. Song, Y. Huang, and H. Lu, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 2395–2400, doi: https://doi.org/10.1109/SMC52423.2021.9659106.

[5] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, Feb. 2019, doi: http://arxiv.org/abs/1902.06673.

[6] H. T. Phan, N. T. Nguyen, and D. Hwang, "Fake news detection: A survey of graph neural network methods," *Appl. Soft Comput.*, vol. 139, Art. no. 110235, June 2023, doi: https://doi.org/10.1016/j.asoc.2023.110235.

[7] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, "Evidence-aware fake news

detection with graph neural networks," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2501–2510, doi: https://doi.org/10.1145/3485447.3512122.

[8] J. Zhu, C. Gao, Z. Yin, X. Li, and J. Kurths, "Propagation structure-aware graph transformer for robust and interpretable fake news detection," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2024, pp. 4652–4663, doi: https://doi.org/10.1145/3637528.3672024.

[9] J. Xu, J. Chen, S. You, Z. Xiao, Y. Yang, and J. Lu, "Robustness of deep learning models on graphs: A survey," *AI Open*, vol. 2, pp. 69–78, June 2021, doi: https://doi.org/10.1016/j.aiopen.2021.05.002.

[10] Y. Liu and Y. F. Wu, "Early detection of fake news on social media through propagation path classification," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2018, pp. 354–361.

[11] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proc. Int. AAAI Conf. Web Soc. Media*, June 2020, pp. 626–637, doi: https://doi.org/10.1609/icwsm.v14i1.7329.

[12] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, p. 1348, June 2021, doi: https://doi.org/10.3390/electronics10111348.

[13] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. ACM Int. Conf. Multimedia (MM)*, Oct. 2019, pp. 1942–1951, doi: https://doi.org/10.1145/3343031.3350850.

[14] P. Wang, H. Shi, X. Wu, and L. Jiao, "Sentiment analysis of rumor spread amid COVID-19: Based on Weibo text," *Healthcare*, vol. 9, no. 10, p. 1275, Oct. 2021, doi: https://doi.org/10.3390/healthcare9101275.

[15] H. Khazane, M. Ridouani, F. Salahdine, and N. Kaabouch, "A holistic review of machine learning adversarial attacks in IoT networks," *Future Internet*, vol. 16, no. 1, Art. no. 32, Jan. 2024, doi: https://doi.org/10.3390/fi16010032.

[16] L. Hu *et al.*, "Dual-aspect active learning with domain-adversarial training for low-resource misinformation detection," *Mathematics*, vol. 13, no. 11, Art. no. 1752, June 2025, doi: https://doi.org/10.3390/math13111752.