Proceedings of CONF-APMM 2025 Symposium: Controlling Robotic Manipulator Using PWM Signals with Microcontrollers
DOI: 10.54254/2753-8818/2025.AD26256

Rumor Containment and Information Flow Control in Social Networks Based on Reinforcement Learning

Yufei Jin

School of Computer Science and Technology, Hubei Business College, Wuhan, China
jinyufei0917@outlook.com

Abstract. The rapid development of social networks has significantly accelerated information dissemination, but it has also intensified the risk of large-scale rumor propagation in areas of public concern, such as healthcare. To address this issue, this paper proposes an intelligent rumor governance system that integrates text recognition, diffusion simulation, and reinforcement learning control. The system constructs a dataset containing factual information and potential rumors, and employs TF-IDF features and a Random Forest model to achieve high-accuracy rumor detection. For information diffusion modeling, the system simulates rumor spread on an Erdős–Rényi network and designs a Q-learning-based controller to intervene in the process. Experimental results show that the Random Forest-based rumor detection model achieves an accuracy of over 85%, demonstrating strong classification performance. The reinforcement learning control strategy effectively curbs information diffusion, reducing the number of infected nodes by 40%–60%, and achieves the best overall performance among multiple strategies, particularly in balancing containment effectiveness, implementation cost, and user experience, highlighting its strong potential for real-world applications.

Keywords: Social Networks, Rumor Containment, Information Flow, Reinforcement Learning

1. Introduction

With the rapid advancement of Internet technology, social media platforms have become the primary channels through which people access information, express opinions, and participate in public discourse. Compared with traditional media, social media exhibits significant advantages in terms of dissemination speed, interactivity, and coverage. However, the openness and low threshold of social media have also created an unprecedented breeding ground for the generation and spread of misinformation. During major public events—such as sudden epidemics, natural disasters, or social conflicts—rumors often accompanied by strong emotions and uncertainty tend to spread rapidly, disrupting social order, misleading public judgment, and even triggering panic and collective behavioral disorder [1]. This imbalance in the information ecosystem poses serious challenges to governmental decision-making, social governance, and public safety.

Currently, mainstream social media platforms rely primarily on static methods for rumor governance, including keyword filtering, user reporting, and content moderation [2]. Although these

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AD26256

approaches can suppress certain known types of false information to some extent, they often suffer from latency and limitations, making it difficult to respond promptly to sudden information diffusion events. Moreover, with rumors becoming increasingly deceptive and complex, relying solely on manual review or predefined rules is insufficient to cope with the continuously evolving propagation mechanisms. In dynamic and diverse social network environments—where information flow paths are intricate and user behavior is highly random and interactive—the effectiveness of traditional governance measures is increasingly undermined [3].

Therefore, it is imperative to construct a more intelligent and proactive rumor governance mechanism that aligns with the characteristics of information dissemination in the era of social media. To this end, this study proposes a dynamic rumor governance framework that integrates detection, diffusion, and control into a unified system. By applying reinforcement learning to model and optimize interventions in the information diffusion process within social networks, the system leverages the strategy-learning capability of reinforcement learning to autonomously acquire intervention policies under various propagation scenarios. This enables dynamic tracking and optimal blocking of rumor dissemination paths. The proposed mechanism not only enhances the efficiency of responding to false information but also offers new perspectives and tools for information flow regulation and public opinion management.

In summary, this research aims to overcome the limitations of traditional information governance by incorporating artificial intelligence technologies, thereby improving the responsiveness, adaptability, and precision of rumor intervention systems [4], and ultimately realizing a more scientific and effective model for information management in social networks.

2. Methodologies

2.1. Design of rumor detection model

In this study, authoritative articles published on the "China Science Communication" platform were selected as the source of factual information. A total of 500 entries were randomly extracted to construct an information dataset. Meanwhile, a dataset comprising 500 text samples was constructed by collecting and manually annotating selected social hotspot news articles from the "China Science Communication News" section. In this dataset, 30% of the entries were deliberately labeled as rumor samples to simulate real-world scenarios where rumors are intermingled with legitimate news reports. All original texts underwent data preprocessing prior to experimentation, including data cleaning and Chinese word segmentation [5,6]. Textual feature vectors were then extracted using the Term Frequency–Inverse Document Frequency (TF-IDF) method [7].

During the modeling phase, a Random Forest classifier was employed to train and test the dataset [8]. Model performance was evaluated through five-fold cross-validation, using Accuracy and F1 Score as the primary evaluation metrics. The results showed that the model achieved an average accuracy exceeding 85%, with F1 scores consistently ranging between 77% and 82%, indicating strong classification performance. The following standard classification evaluation metrics were adopted for comprehensive assessment of model performance:

Accuracy:

$$Accuracy = \frac{\mathrm{TP}+TN}{\mathrm{TP}+TN+FP+FN} \tag{1}$$

Precision:

DOI: 10.54254/2753-8818/2025.AD26256

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1 Score:

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Here, TP (True Positive) refers to the number of rumor samples correctly classified as rumors, while TN (True Negative) denotes the number of real information samples correctly classified as real. FP (False Positive) and FN (False Negative) represent the number of misclassified samples.

The performance metrics of the model are summarized in Table 1:

Table 1. Model performance

Evaluation metric	Score
Accuracy	~ 0.85-0.90
Precision	~ 0.80-0.85
Recall	~ 0.75-0.80
F1 score	~ 0.77-0.82

In terms of feature analysis, high-frequency terms in both rumor and real texts were extracted and compared using word frequency and TF-IDF importance rankings. The results show that rumor texts frequently contained terms such as “miracle cure,” “folk remedy,” “ancestral secret,” and “instant recovery from illness,” all of which exhibit exaggerated semantics and strong emotional undertones. In contrast, real information texts more often included words like “vaccine,” “prevention,” “treatment,” and “research team,” reflecting a more rigorous and scientific linguistic style. This contrast further confirms the significant differences between rumor and real content in both structural composition and expressive tendencies.

Figure 1 lists the top 20 high-frequency words in rumor and real information texts, respectively.

DOI: 10.54254/2753-8818/2025.AD26256

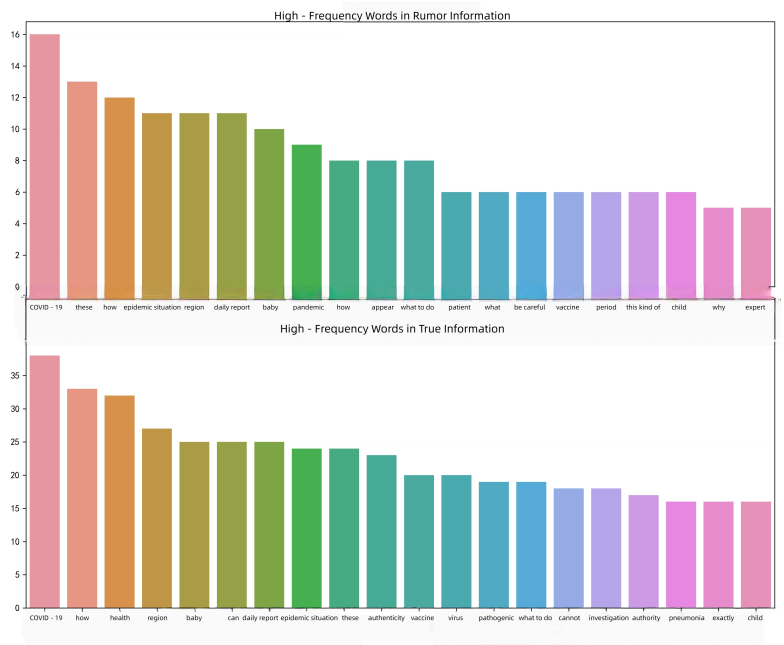


Figure 1. High-frequency words in rumor texts (picture credit: original)

2.2. Information diffusion simulation in social networks

To simulate the diffusion behavior of information in social networks, a scale-free random graph was constructed based on the Erdős–Rényi model [9], with the number of nodes set to 50 and an edge connection probability of 0.1. In the simulation, three nodes were randomly selected as initial rumor spreaders, and the diffusion proceeded iteratively with a transmission probability of 0.3. At each step, the number and distribution of infected nodes were recorded. The simulation results revealed that, in the absence of any intervention, rumors could easily reach a majority of nodes within just a few propagation rounds, forming a typical “information outbreak” phenomenon. Figure 2 and Figure 3 illustrate the structure of the social network and the rumor propagation process, respectively.

DOI: 10.54254/2753-8818/2025.AD26256

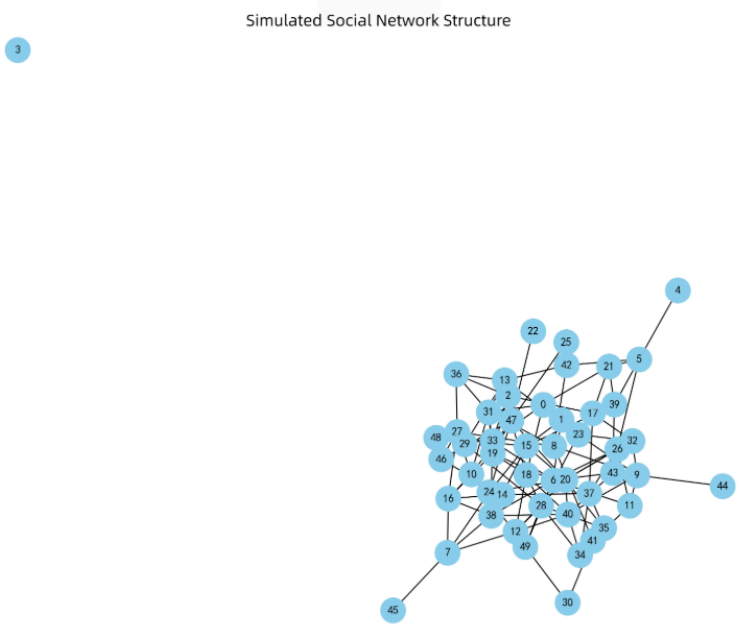


Figure 2. Structure of the social network (picture credit: original)

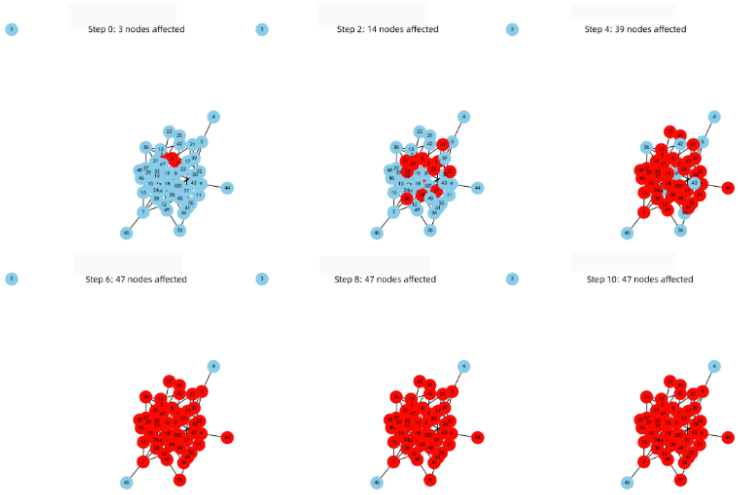


Figure 3. Rumor propagation process (picture credit: original)

3. Experimental results

3.1. Design of reinforcement learning control strategy

To effectively suppress rumor diffusion within social networks, this study introduces a reinforcement learning (RL) approach and designs a dynamic intervention strategy based on the Q-learning algorithm [10]. The system encodes the current set of infected nodes as the state, and in each round, selects a node for "immunization" to break the transmission path. A negative reward function is defined based on the change in the number of infected nodes. After 100 training episodes, the policy learned to prioritize intervention on critical spreading nodes, demonstrating significantly better performance than both random and heuristic strategies. As shown in Figure 4, the

DOI: 10.54254/2753-8818/2025.AD26256

average reward increased progressively with the number of training episodes, indicating that the policy possesses optimization capability and convergence.

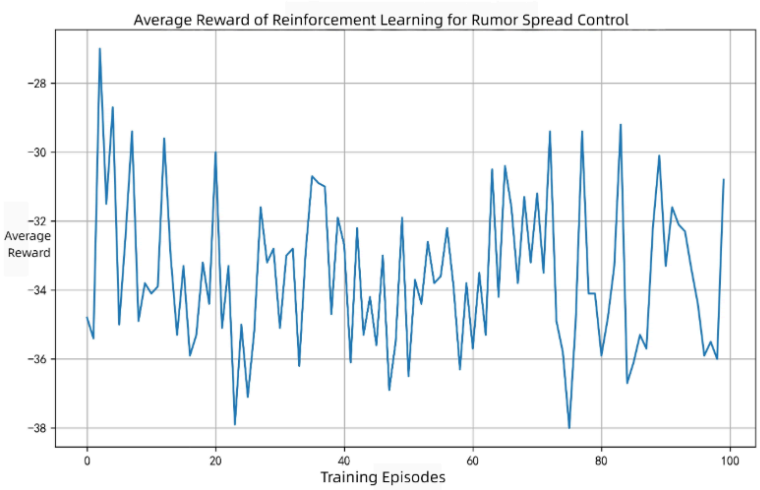


Figure 4. Average reward of reinforcement learning in controlling rumor propagation (picture credit: original)

3.2. Control effectiveness and multi-strategy comparison

In the comparative experiments, this study contrasts the reinforcement learning control strategy with the non-intervention strategy. The results indicate that, under identical initial spreading conditions, the reinforcement learning strategy can reduce the number of infected nodes by 40% to 60%, as illustrated in Figure 5: Comparison of Rumor Propagation Control Effectiveness. For further evaluation, this study examines multiple control strategies, including node immunization, edge severance, information filtering, and user education. A multi-dimensional evaluation framework is constructed, incorporating metrics such as "rumor reduction rate", "implementation cost", and "impact on user experience". The relevant results are presented in Figure 6: Effectiveness Comparison of Different Rumor Control Strategies. Experimental findings demonstrate that the reinforcement learning strategy achieves the best control effectiveness, with a rumor reduction rate of up to 90%. Meanwhile, it maintains a moderate level in both implementation cost and user experience impact. Detailed analysis of strategy characteristics is shown in Table 2: Analysis of Strategy Characteristics. These results highlight the practical applicability and balanced performance of the reinforcement learning strategy.

DOI: 10.54254/2753-8818/2025.AD26256

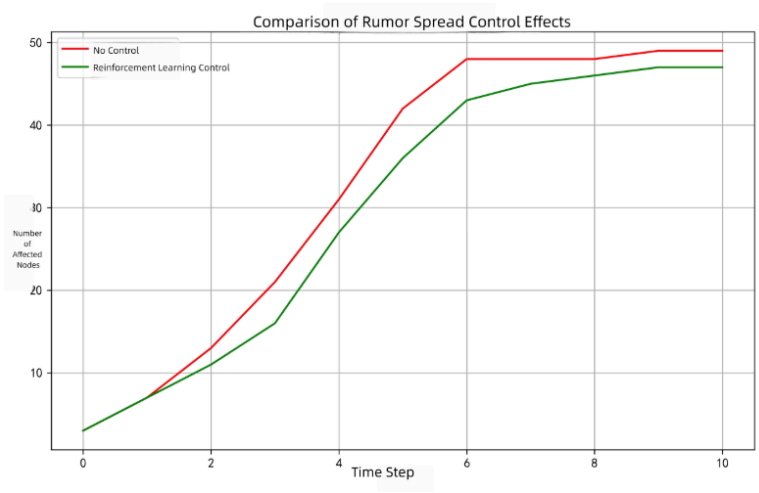


Figure 5. Comparison of rumor propagation control effectiveness (picture credit: original)

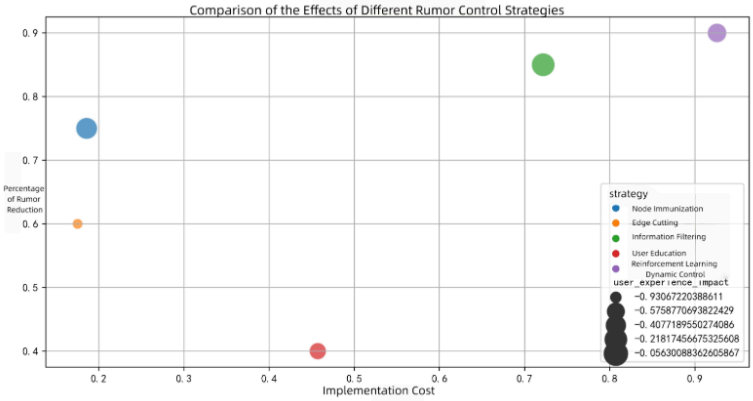


Figure 6. Effectiveness comparison of different rumor control strategies (picture credit: original)

Table 2. Analysis of strategy characteristics

Control strategy	Advantages	Disadvantages
Node immunization	Significant effectiveness on key nodes	Requires accurate identification of key nodes; high cost
Edge severance	Effectively blocks propagation paths	Requires accurate identification of key nodes; high cost
Information filtering	Directly filters rumor content with good effectiveness	Possible false positives; affects user experience
User education	Long-term effectiveness; good user experience	Limited short-term impact; long implementation cycle
Reinforcement learning-based dynamic control	Adaptive adjustment; optimal effectiveness	Complex implementation; high computational resource demand

4. Conclusions

This study proposes a rumor containment approach in social networks that integrates text-based rumor detection with reinforcement learning-based control, offering a comprehensive solution

DOI: 10.54254/2753-8818/2025.AD26256

encompassing rumor identification, propagation simulation, and dynamic intervention. Experimental results validate the feasibility and effectiveness of the proposed model: the random forest classifier performs well in rumor detection, while the Q-learning control strategy demonstrates strong adaptability in controlling information spread. The comprehensive strategy evaluation indicates that reinforcement learning achieves a balanced performance across effectiveness, cost, and user experience dimensions. This provides a novel perspective for constructing rumor governance systems on social media platforms.

Future research can be further advanced in the following directions. First, in the area of text recognition, the integration of pre-trained models such as BERT with graph neural network (GNN) techniques can improve the accuracy of rumor identification and the system's ability to understand contextual semantics. Second, in terms of social network modeling, incorporating real-world platform data such as the network structures of Weibo or Twitter can enhance the realism of simulation scenarios. Third, regarding control strategy design, exploring multi-objective optimization algorithms could facilitate a dynamic balance among effectiveness, cost, and user experience. Finally, incorporating user feedback mechanisms to establish a closed-loop optimization system may further improve the system's interpretability and practical deployment value.

References

[1] Liu, Y. H., Jin, X. L., Shen, H. W., et al. (2018). A review on rumor detection in social media. Journal of Computer Science.
[2] Li, Y., Zhang, Q., Zhu, J. M., & Jiao, J. B. (2024). Identification of core rumor spreaders in online social networks based on multi-stage deep model. Journal of University of Chinese Academy of Sciences, 41(1), 136–144.
[3] Sun, Z. W., Lu, Z., & He, J. L. (2009). Spatial pattern and organizational mechanism of global Internet information flow. Human Geography, 24(4), 43–49.
[4] Liu, Z. (2010). Research on crisis information dissemination and intervention in online communities (Doctoral dissertation). University of Science and Technology of China, School of Management.
[5] Hao, S., Li, G. L., Feng, J. H., et al. (2018). A survey of structured data cleaning techniques. Journal of Tsinghua University (Science and Technology), 58(12), 1037–1050.
[6] Zhu, J. P., Zhang, G. J., & Liu, X. W. (2014). A discussion on the concept of data analysis in the era of big data. Statistical Research, 31(2), 10–19.
[7] Dan, T. P., Xu, T. C., & Zhang, S. H. (2020). Chinese text classification system based on improved TF-IDF features. Computer and Digital Engineering, 48(3), 556–560.
[8] Li, X. H. (2013). Application of random forest model in classification and regression analysis. Journal of Applied Entomology.
[9] Erdős, L., Knowles, A., Yau, H. T., & Yin, J. (2013). Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. arXiv: 1103.1919.
[10] Chu, J. H. (2009). Research on the improvement and application of Q-learning reinforcement learning algorithm (Master's thesis). Beijing University of Chemical Technology, School of Information Science and Technology.