Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 13, No. 2, 2025

A Review of Intelligent Public Opinion Governance in Security and Protection Engineering

Yichi Zhang, Fanliang Bu *

College of Information Network Security, People's Public Security University of China, Beijing, China
*** Corresponding author:** Fanliang Bu (Email: bufanliang@sina.com)

Abstract: This paper systematically explores the research methods and practical frameworks of intelligent public opinion governance in the context of social media, centered on the theory of public opinion governance within the secondary discipline of Security and Protection Engineering. Firstly, it analyzes the theoretical connotations of public opinion governance and its role in social risk prevention and control, clarifying the role and challenges faced by public security organs in public issue governance. Secondly, it constructs a fundamental framework for intelligent public opinion governance covering pre-event warning, early identification, situation prediction, control intervention, and fact-checking, providing a systematic and structured methodological foundation for public opinion governance. Based on this framework, and focusing on the disciplinary hotspot of intelligent public opinion governance on social media, the paper provides an in-depth review of recent advancements in related frontier research. Representative technical methods and solution pathways are systematically summarized according to the five governance stages, including: pre-event warning based on large language models, early identification through multimodal signal fusion, situation prediction using temporal graph neural networks, control intervention driven by reinforcement learning, and fact-checking combining retrieval-augmented generation with causal inference. Through this systematic review, this paper aims to provide theoretical references, methodological support, and technical insights for research on intelligent public opinion governance under the discipline of Security and Protection Engineering, promoting a paradigm shift in public opinion governance from experience-driven to data-driven, model-driven, and intelligent decision-making.

Keywords: Security and Protection Engineering; Social Media; Intelligent Public Opinion Governance; Deep Learning.

1. Introduction

In the current information ecosystem centered on social media, the formation, diffusion, and evolution of public issues exhibit characteristics of real-time nature, high speed, and decentralization. The diversification of information dissemination channels and the significant lowering of barriers to user participation enable the public to form attitudes and engage in emotional expression as soon as signs of an event emerge, presenting unprecedented challenges for public governance. Particularly in sudden incidents, social risk hazards, and other issues of collective concern, the trajectory of public opinion often carries high uncertainty, and its spread far outpaces traditional government response mechanisms. Consequently, public opinion events are increasingly becoming significant risk sources affecting social order and public safety.

Social media public opinion events typically refer to the collective expression process of opinions, emotions, and attitudes spontaneously generated, widely disseminated, and continuously fermented by users on social platforms around specific public events, social issues, or group appeals. Such events are characterized by rapid aggregation, complex diffusion chains, and strong heterogeneity among dissemination agents, and are more prone to explosive growth in highly attention-grabbing or sensitive topics. Once public opinion spirals out of control, it may lead to social panic, group polarization, the spread of false information, misunderstanding and erosion of trust in public institutions, and even induce associated risks in the real world. Therefore, effectively identifying and managing public opinion risks has become a crucial component of modernizing national governance systems.

Within China's social governance system, public security organs bear the important responsibilities of maintaining social stability, resolving social conflicts, and preventing risk escalation. As public expression spaces and arenas for public discussion increasingly migrate online, social media public opinion governance has become a critical component for public security organs in risk warning, emergency management, and social stability maintenance. Public security public opinion governance is not only related to the capacity to respond to emergencies but also concerns how to achieve timely information acquisition, accurate analysis, and reasonable intervention in environments where objective facts are unclear, public sentiment is highly sensitive, and misinformation and biased information are prone to overlap. Its primary tasks include: identifying potential risks in advance to reduce the accumulation of social conflicts; accurately grasping the structure of public attitudes to enable targeted interventions; enhancing the effectiveness of information dissemination by public institutions and strengthening communication links with the public; and establishing a standardized, scientific intelligent public opinion governance system to improve overall social governance efficacy.

Against this backdrop, constructing a systematic and operable intelligent public opinion governance system has become an inevitable requirement for maintaining social stability and public order. Intelligent public opinion governance is a systematic engineering endeavor encompassing prediction, monitoring, modeling, and intervention, involving stages such as pre-event warning, early identification, situation prediction, control intervention, and fact-checking. These stages together form a complete governance chain from the prediction and identification of

public opinion to the execution of interventions, providing a theoretical foundation and practical basis for the timely response, precise intervention, and scientific assessment of public opinion. They also offer a structured perspective for understanding the formation mechanisms and intervention logic of public opinion risks. This paper systematically reviews the current technical methods and research progress in the field of intelligent public opinion governance around this chain, aiming to establish a systematic analytical framework and provide insightful ideas for subsequent related work.

2. The Fundamental Framework of Intelligent Public Opinion Governance

The core objective of intelligent public opinion governance is to provide scientific support for the entire lifecycle of a public opinion event—from germination and diffusion to intervention—through data-driven approaches and model inference, thereby shifting governance actions from passive response to active perception and dynamic regulation. In the process of managing social media, intelligent public opinion governance is typically constructed as a continuous technical framework revolving around pre-event warning, early identification, situation prediction, control intervention, and fact-checking. It emphasizes the synergistic effects of forward-looking perception, dynamic modeling, and strategy optimization, providing a quantifiable, verifiable, and intervenable technical foundation for public safety and social stability.

Intelligent public opinion governance should commence with pre-event warning, focusing on risk perception and trend judgment during the stage before public opinion fully forms. The core of this stage is not identifying specific events but rather capturing potential triggers and aggregation trends of public opinion through continuous modeling of user stances, emotional tendencies, and opinion distribution structures regarding specific topics on social media. At the technical level, this can rely on methods such as user stance recognition, attitude polarization analysis, and group opinion structure modeling. By delineating the size evolution, intensity of opposition, and interaction patterns of groups with different stances, it provides a basis for early analysis of public opinion risks. The key to pre-event warning lies in possessing the alerting capability for "potential public opinion," treating public opinion as a group phenomenon formed by the gradual aggregation of individual attitudes, thereby making public opinion risks somewhat predictable.

As public opinion forms and evolves into its initial stage, the early identification phase focuses on how to accurately capture event-level occurrences while public opinion is still at a low level of heat and weak diffusion. Unlike pre-event warning, which emphasizes macro trends, early identification pays more attention to the emergence of specific events and abnormal changes in their dissemination signals. This stage typically utilizes signals such as topic surge detection, abnormal analysis of dissemination structures, changes in user behavior patterns, or abrupt shifts in emotional expression to identify early-stage public opinion events. Relevant technologies encompass event detection, anomaly detection, weak signal mining, and multimodal information fusion. The challenge lies in effectively distinguishing genuine public opinion events from transient, non-risky information fluctuations within the high-noise, strong-background-interference environment of social media. The effectiveness of early identification directly impacts whether subsequent analysis and intervention can secure a critical time window.

Between the early identification and control intervention of public opinion events, situation prediction plays a bridging role, focusing on deducing and assessing the uncertainty of the future evolution path of a public opinion event. This stage will answer questions such as whether public opinion will continue to ferment, whether there is a risk of losing control, and what kind of critical turning points may emerge. Technically, public opinion is often modeled as a dynamic system, such as a time series, evolving over time. By comprehensively analyzing dissemination scale, emotional trends, structural evolution, and user participation patterns, it predicts the development trend of public opinion within different time windows. The goal of situation prediction is not simple trend extrapolation but to provide actionable decision windows for governance actors, enabling intervention actions to seek an optimal balance between cost and effectiveness.

Subsequently, the control intervention stage translates the aforementioned analysis and predictions into concrete, actionable governance strategies. This stage involves not only the decision of whether to intervene but also the optimization of intervention timing, methods, and intensity. Related research typically revolves around modeling information guidance strategies, assessing intervention effects, and dynamically optimizing strategies, emphasizing effective regulation of the public opinion situation while avoiding triggering secondary risks. Technically, it requires integrating the mechanisms of public opinion dissemination, user behavior response characteristics, and strategy feedback relationships to simulate and compare the potential impacts of different intervention methods, thereby supporting more targeted and differentiated governance decisions.

Fact-checking, as a fundamental task throughout the entire process of intelligent public opinion governance, is key to ensuring the scientific nature of governance decisions and the credibility of the public information environment. False information, one-sided narratives, and emotionalized content often accelerate the spread of public opinion and amplify risks. Therefore, continuous verification of information authenticity is highly significant for decision-making in the warning, early identification, situation prediction, and control intervention stages of public opinion events. Fact-checking methods typically involve assessing the credibility of information sources, verifying content consistency, and comparing evidence from multiple sources. Their results not only directly influence public perception but also provide critical evidence for governance actors in information dissemination and strategy selection.

The fundamental framework of intelligent public opinion governance integrates pre-event warning, early identification, situation prediction, control intervention, and fact-checking into a logically coherent, well-structured technical system. While providing a clear outline for an in-depth review of key technical pathways, representative methods, and research progress at each stage, it also offers a theoretical foundation and solution framework for strengthening the scientific orientation and problem-targeting of the governance process, promoting the precision and intelligence of public opinion governance.

3. Frontier Research on Intelligent Public Opinion Governance

Centered on the aforementioned technical framework for intelligent public opinion governance, researchers in recent years have systematically conducted extensive empirical and methodology-oriented research published in top-tier journals and conferences, propelling the study of public opinion governance issues from an experience-driven approach towards a model-based, data-driven, and verifiable research paradigm. Leveraging large-scale social media data, these studies have proposed various representative modeling ideas and technical pathways for key stages such as pre-event warning, early identification, situation prediction, control intervention, and fact-checking of public opinion risks, validating them on real or quasi-real data. The relevant achievements have not only continuously introduced more complex mechanisms of representation learning, temporal modeling, and decision modeling at the methodological level but have also gradually refined the research boundaries and core objectives of each stage of intelligent public opinion governance in terms of problem definition and evaluation systems. The following sections will systematically review and analyze representative research works published in recent years in top-tier journals and conferences, organized according to the five stages of the fundamental framework for intelligent public opinion governance.

3.1. Pre-event Warning

In the pre-event warning stage, related research focuses on identifying potential triggers and aggregation trends of public opinion through continuous modeling of user stances, emotions, and opinion structures on social media, thereby enabling prediction of potential public opinion events. In recent years, scholars have introduced various advanced computational models and technical pathways to enhance the accuracy and foresight of warnings.

Suh et al. [1] focused on utilizing large language models to predict opinion distributions of specific social subgroups. They fine-tuned language models on a large scale to enable them to directly predict the response probability distribution of a given subgroup towards a specific issue based on demographic and socio-economic descriptions. The emphasis of this work lies in identifying group attitude divergence and polarization trends that may signal public opinion events from the structural characteristics of macro opinion distributions, providing a quantitative tool for pre-event warning from the perspective of social mentality.

Wang et al. [2] approached from the perspective of dynamic system modeling, constructing a GAN-SEIR framework to simulate the dynamic process of public opinion evolution in social networks. This research combined Generative Adversarial Networks (GAN) with the classic epidemic compartmental model (SEIR), using a generator to simulate the diffusion of public opinion information and a discriminator to assess its authenticity, thereby modeling the complex conversion process of users from exposure to information, to belief, and then to dissemination. This model aims to deduce the pre-event diffusion and critical turning points through simulation of evolutionary paths, providing decision support based on dynamic simulation for pre-event warning of public opinion events.

Huo et al. [3] proposed a method for predicting public opinion event types by combining Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). This method dynamically retrieves relevant data slices from a knowledge base through a retrieval mechanism combining dense vector matching and the BM25 algorithm, and generates high-quality few-shot prompts to enhance the LLM's ability to classify and predict potential public opinion event types in scenarios with limited labeled data. Its core lies in using retrieved similar historical cases as context, enabling the model to more accurately identify the characteristics of potential public opinion events.

Li et al. [4] addressed a practical task—Target-Stance Extraction (TSE)—aiming to automatically identify target entities from social text and predict attitudes towards them, thereby overcoming the "target-known" assumption limitation in traditional stance detection. The study proposed a two-stage framework: first identifying potential targets in the text through target classification or keyword generation, and then predicting stances towards the targets in a multi-task manner. This method is not only applicable to known target detection but can also be extended to zero-shot scenarios, providing a feasible pathway for stance inference regarding new targets. This research provides important technical support for dynamic stance perception in public opinion warning, enabling systems to effectively capture attitudinal structures even when targets are unclear, thereby enhancing the foresight and coverage of pre-event warning.

Gatto et al. [5] proposed a Chain-of-Thought Embeddings method, which integrates the reasoning process of LLMs as features into traditional stance detection models, mitigating confusion and hallucination issues of LLMs in implicit stance identification. By incorporating CoT reasoning text, this method enhances the model's understanding of implicit stances and domain-specific patterns, achieving state-of-the-art performance on multiple stance detection datasets. This research provides new insights on how warning systems can effectively leverage the reasoning capabilities of LLMs, improving robustness in stance identification within dynamic and implicit expression scenarios.

Li et al. [6] proposed a Knowledge-Augmented Stance Detection framework (KASD), which significantly improved the accuracy and generalization ability of stance detection by integrating event knowledge retrieved from Wikipedia and discourse knowledge generated by ChatGPT. The research constructed heuristic retrieval and LLM filtering mechanisms, enabling the model to better understand implicit stance expressions, particularly excelling in zero-shot scenarios. This work propels pre-event warning of public opinion from pure text analysis towards knowledge-enhanced deep semantic understanding, providing a more reliable technical pathway for stance capture and pre-event trend analysis in complex contexts.

Zhang et al. [7] addressed the resource scarcity issue in cross-lingual stance detection by proposing an LLM-enabled Knowledge Elicitation and Retrieval framework (KEAR). By extracting background knowledge, reasoning knowledge, and explanatory knowledge from the LLM's reasoning process and constructing a hierarchical knowledge retrieval mechanism, they significantly improved the performance of zero-shot cross-lingual stance detection. This research not only overcomes semantic gaps between languages but also provides a feasible technical solution for cross-lingual stance perception in multilingual public opinion monitoring, enhancing the coverage and adaptability of warning systems on global issues.

These seven studies have deepened the theory and methods of the pre-event warning stage from technical dimensions such as group opinion distribution prediction, information diffusion dynamic simulation, dynamic stance perception, and cross-lingual stance perception. They promote the evolution of public opinion warning from traditional monitoring based on explicit signals towards a forward-looking and intelligent perception stage based on deep semantic understanding, group psychology inference, complex system simulation, dynamic stance perception, and cross-lingual knowledge fusion. They provide diversified, multi-layered technical support for constructing a scientific early warning system for public opinion risks.

3.2. Early Identification

In the early identification stage of intelligent public opinion governance, related research focuses on rapidly and accurately capturing signals of sudden events that have not yet diffused on a large scale from the high-noise, highly dynamic dissemination environment of social media, identifying the emergence of specific public opinion events and early abnormal changes in their propagation dynamics. This stage emphasizes the real-time detection and fusion analysis of early signals from multiple dimensions such as topic surges, abnormal dissemination structures, and abrupt changes in user behavior patterns, aiming to secure a critical time window for subsequent analysis and intervention. In recent years, scholars have progressively improved the precision and timeliness of early identification by combining technical pathways such as event detection, anomaly analysis, and multimodal signal fusion.

El-Mefleh et al. [8], although not directly analyzing social media content in their constructed Social Early Warning System (SEWS), proposed a standardized social unrest index that provides crucial input for the early identification of the macro context for public opinion governance. By monitoring abnormal fluctuations in macroeconomic indicators such as oil prices and unemployment rates, this research aims to identify socio-economic stressors and macro risk precursors that may trigger large-scale social media public opinion. While its model has limitations in capturing specific, event-level early micro-signals on social media, it provides early, structured warning indicators from the socio-economic dimension for analysts to anticipate the macro-issue sentiments and discussion foci that may surge onto social platforms, thereby moving the starting point of public opinion identification from the online event itself forward to its potential offline macro triggers.

Kim et al. [9], using discussions on Universal Basic Income (UBI) on Reddit as a case study, proposed a content-user-community triple-driven counterfactual modeling framework to analyze the micro-level driving mechanisms behind macro stance changes. Through techniques like topic extraction, user cohort analysis, and community embedding, this research distinguishes whether changes in user stances are caused by distribution shifts in discussion topics, participating populations, or discussion settings, or by the intrinsic transformation of stances within groups. This method can identify early-stage group attitude differentiation and issue focusing processes during the formation of public opinion from large-scale online conversations, providing a fine-grained analysis tool for the early emotional structure and issue evolution of public opinion events.

Parekh et al. [10] constructed the SPEED framework for epidemic event early warning directly targeting social media data, with the core objective of achieving early detection ofpublic health emergency. This research innovatively defined seven types of generic epidemic event ontologies and trained models to detect surges in mentions of these events from tweet streams in real-time. This method bypasses reliance on specific disease names or sentiments; by monitoring the abnormal aggregation of event-level discussion signals, it can capture the early emergence patterns of structured event discussions in social media before epidemic-related public opinion erupts on a large scale. This provides a scalable event detection technical paradigm for the automated, real-time identification of early topic formation and diffusion signals for public opinion in public emergencies from the vast sea of social noise.

Shang et al. [11] proposed a framework based on swarm intelligence and domain-adaptive graph learning for early detection of emerging misinformation in the healthcare domain. Addressing the challenges of lacking labeled data and medical knowledge for emerging health topics, they designed a knowledge-driven domain adaptation mechanism. By constructing a medical knowledge information network and incorporating crowd-sourced expert knowledge to update and correct knowledge triplets, the framework achieves accurate early identification of health misinformation in the target domain. This framework not only improves early detection accuracy but also provides a cross-domain knowledge adaptation pathway for knowledge transfer and signal enhancement in early identification of public opinion events across multiple domains.

Hu et al. [12], from the perspective of multimodal fake news early detection, proposed a Multimodal Prompt Learning framework (MPL) based on CLIP. Leveraging bimodal information from images and text, and through learnable prompt vectors and a multimodal feature fusion module, this research enhances the semantic understanding and classification capability for early signals of fake news. MPL maintains high detection performance even with limited labeled data, providing an effective lightweight technical solution for multimodal information fusion in early identification of public opinion events and signal extraction in low-resource environments.

Martín-Corral et al. [13] further explored, from the perspective of social media sensors, how to use high-centrality users on Twitter as early signal sensors to detect outbreaks of influenza-like illnesses. By analyzing features such as user out-degree, mobility patterns, and content topics, this research identified user groups that publish related information earlier before an outbreak and verified the effectiveness and lead time of these sensors in early epidemic warning. This method provides empirical support for utilizing user network positions and behavioral characteristics to identify early social sensing signals forpublic health emergency, highlighting the importance of integrating user network topology and behavioral characteristics in the early identification of public opinion.

These studies construct a multidimensional technical framework for the "early identification" stage in social media intelligent public opinion governance from dimensions such as macro risk precursor warning, social media event surge detection, cross-domain information early identification, and key user behavior sensing. The results indicate that effective early identification mechanisms have moved beyond relying on single heat metrics, now requiring the integration of

composite technical pathways such as macro-cause analysis, micro-behavior sensing, multimodal content analysis, cross-domain knowledge transfer, and event-level signal capture. This comprehensive system establishes a solid methodological foundation and indicates feasible practical directions for constructing an intelligent public opinion early identification system with real-time response, precise warning, and proactive intervention capabilities within the dynamic and complex social media ecosystem.

3.3. Situation Prediction

Within the chain of intelligent public opinion governance, situation prediction plays the crucial bridging role from early identification to precise intervention. This stage focuses on deducing and assessing the uncertainty of the future evolutionary path of a public opinion event, aiming to answer core questions such as whether the public opinion will continue to ferment, whether there is a risk of losing control, and what kind of critical turning points may emerge. Technically, public opinion is often modeled as a dynamic system, such as a time series. By comprehensively analyzing dissemination scale, emotional trends, structural evolution, and user participation patterns, it predicts the development trends of public opinion within different time windows. The goal of situation prediction is not simple trend extrapolation but to provide actionable decision windows for governance actors, enabling intervention actions to seek an optimal balance between cost and effectiveness. In recent years, related research has made significant progress in improving the accuracy, dynamism, and interpretability of situation prediction.

Feng et al. [14] proposed a Sphere-effect based Information diffusion prediction model on Large-scale social Networks (SILN). This research broke through the limitation of homogenizing participant influence in traditional sequential modeling by introducing the sphere effect from dual perspectives of structure and time for the first time, differentiating the varying influence of different participants during the propagation process. By efficiently extracting subgraphs relevant to the current propagation and combining optimized graph storage techniques, the model significantly reduces computational and storage overhead, enabling real-time prediction on million-node-level social networks. SILN can capture structural proximity and temporal phase similarity during propagation, thereby more accurately deducing the diffusion path and key influential nodes of public opinion, providing a scalable solution for situation prediction oriented towards large-scale dynamic networks.

Wang et al. [15], targeting cyberbullying scenarios, proposed a data-driven agent-based Model for Public Opinion Propagation Simulation (MPOPS). This research constructed a public opinion propagation environment incorporating opinion fusion and polarization, and depicted the behavioral characteristics of agents (users) through multi-level fine-grained modeling. The model improved the classic SEIR propagation model by incorporating opinion fusion and polarization, local popularity, and topic interest as key factors influencing propagation probability, thereby enabling the simulation of public opinion evolution trends in cyberbullying events. Through data-driven simulation of the real case of the 2022 Tangshan barbecue restaurant assault incident, this work validated the model's effectiveness in predicting public opinion propagation trajectories and heat evolution, providing simulation-based decision support for situation deduction and intervention point identification in specific negative public opinion scenarios.

Donkers et al. [16], to explore user perception and behavior patterns in online polarized environments, constructed a human-agent interaction synthetic social network experimental framework based on LLM agents. By controlling the opinion distribution of agent groups and recommendation system biases, this research simulated polarized versus moderate discussion atmospheres from the early stages and introduced human participants for natural interaction. Experiments found that polarized environments significantly enhanced users' perception of discussion emotional intensity and group identity salience while reducing expressions of uncertainty. This work not only provides a controlled experimental platform for causal inference regarding online polarization mechanisms but also offers a simulation and empirical research method based on synthetic social environments for assessing and predicting the impact of discussion atmospheres on user participation and opinion formation after early identification of public opinion.

Cisneros-Velarde [17] focused on the opinion dynamics and bias evolution mechanisms in multi-agent systems involving LLMs, systematically analyzing the impact of fairness-consensus bias, caution bias, and safety bias exhibited by LLM agents during interaction on opinion evolution, using funding allocation as a case study. The research found that even on negative issues, the tension between safety bias and consensus bias could still lead to the persistence of funding opinions, and the form of opinion expression significantly affects the diversity of the final opinion distribution. This research reveals potential risks of opinion solidification and value conflicts in LLM group interactions, providing important theoretical and experimental basis for predicting opinion evolution paths and identifying systemic biases in multi-agent collaboration or automated decision-making scenarios after the early identification of public opinion.

Zhong et al. [18], addressing the issues of cross-cascade dependency and cross-platform robustness insufficiency in information diffusion prediction tasks, proposed a Cascade-Retrieved in-context learning framework (CARE). Drawing on the in-context learning idea from LLMs, this framework constructs a dynamic prompt pool from historical cascades and selects the most relevant diffusion patterns as context input based on a retrieval mechanism. CARE further introduces prompt enhancement strategies such as user masking and re-ranking, and combines a social relation embedding module to enhance the completeness of user representation and the robustness of context modeling. Experiments show that this framework exhibits superior prediction performance and stability across cross-platform datasets, effectively identifying similar diffusion patterns from early observed cascade segments, providing high-generalizability prediction support for diffusion trajectory deduction after early identification of public opinion.

Jin et al. [19], targeting popularity trend prediction tasks in social networks, proposed a multi-layer temporal graph neural network framework for learning entity representations and predicting their future trends in heterogeneous and temporal social media environments. Building upon early identification, this research further models public opinion propagation as a sequence of discrete temporal graph snapshots. By introducing a graph structure learning module, node aggregation module, temporal unit, and multi-head attention

layer, it explicitly models the information diffusion effects and temporal evolution dynamics of multi-type relationships between entities. Experimental results indicate that this framework outperforms traditional linear regression, temporal models, and existing heterogeneous graph neural network methods on real datasets including YouTube live streams, MOOC courses, Reddit posts, and Wikipedia edits, significantly improving prediction accuracy under complex relationships and temporal dependencies. This method provides a scalable learning framework for dynamic deduction of popularity trends after early identification of public opinion, balancing structural heterogeneity and temporal evolution characteristics.

The aforementioned studies respectively enhance the efficiency, scalability, and scenario adaptability of prediction models from dimensions such as efficient diffusion prediction on large-scale networks, agent-based simulation for propagation evolution modeling, and evolution mechanism analysis in multi-agent systems. By introducing methods such as controlled experiments, cross-cascade retrieval, multi-agent interaction simulation, and heterogeneous temporal graph modeling, they enhance the dynamism, interpretability, and causal inference capability of public opinion evolution deduction. Together, they promote the evolution of public opinion situation prediction from traditional temporal extrapolation or static analysis towards a deep deduction stage based on dynamic network structures, multi-agent interactions, and graph neural networks. They provide crucial theoretical support and technical tools for constructing a real-time, precise, interpretable, and mechanism-insightful public opinion evolution prediction system.

3.4. Control Intervention

In the control intervention stage of intelligent public opinion governance, related research focuses on how to translate the results of public opinion analysis and prediction into concrete, actionable governance strategies, emphasizing the optimization of intervention timing, methods, and intensity to achieve effective regulation of the public opinion situation while avoiding triggering secondary risks. In recent years, scholars have conducted research around modeling information guidance strategies, assessing intervention effects, and dynamically optimizing strategies. By integrating mechanisms of public opinion dissemination, user behavior response characteristics, and strategy feedback relationships, and through simulating and comparing the impacts of different intervention methods, they aim to support more targeted and differentiated governance decisions.

Wang et al. [20], addressing the opinion maximization problem in social networks, proposed optimizing overall trend of public opinion by modifying the internal opinions of key nodes. Based on the Friedkin-Johnsen (FJ) opinion dynamics model, this research introduced "structural centrality" to quantify a node's potential influence on overall opinion and proposed three efficient algorithms: two approximation methods based on random walk and forest sampling, and an exact selection algorithm based on asynchronous updating. Through local residual propagation and progressive refinement, this algorithm can efficiently and precisely identify the most influential set of nodes in networks with tens of millions of nodes, providing a scalable computational framework for targeted adjustment of key user opinions. This represents an intervention strategy based on modifying internal node opinions, illustrating the technical pathway for implementing precise and efficient interventions in complex networks.

Chu [21], from the perspective of dynamic response and strategy optimization, proposed a Deep Reinforcement Learning-based Crisis Response Optimization framework (DRL-CRO). This framework models the public opinion crisis environment as a dynamic system containing state variables such as public sentiment, information dissemination, and platform activity. The DRL agent selects actions in real-time, including content type, tone, platform, and response timing, to learn strategies aimed at maximizing trust recovery and rumor suppression. This research emphasizes adaptive, real-time, multi-dimensional intervention strategy generation and optimization, breaking through the limitations of traditional static templates and manual decision-making, providing an end-to-end closed-loop optimization solution for implementing dynamic, intelligent communication interventions in evolving public opinion crises.

Ghosh et al. [22], targeting rumor propagation on social media, proposed a dynamic control system based on temporal assessment and delayed intervention. Integrating temporal graph networks and recurrent neural networks, this research performs real-time rumor scoring on tweets and introduces a "temporal embargo" strategy: when model confidence falls within an uncertain interval, the judgment result is delayed, waiting for more contextual information to improve judgment accuracy. By combining dynamic scoring, active learning, dual-model collaboration, and delayed intervention, this system achieves early identification and controllable suppression of rumor propagation. It embodies an intervention approach that regulates information visibility through temporal strategies in rapidly diffusing information environments, representing a confidence-driven temporal intervention mechanism.

Muppasani et al. [23], addressing false information propagation in dynamic opinion networks, proposed an intervention planning framework combining supervised learning and reinforcement learning. In modeling, they introduced continuous opinion and trust representations to enhance realism and developed a node identification method based on ranking algorithms to generate training labels for supervised learning. For large-scale networks, they further proposed a centralized dynamic planner based on a deep value network. Through systematic analysis of different reward functions, it optimizes node selection strategies to maximize intervention effectiveness. This framework emphasizes achieving scalable, adaptive intervention decisions under varying scenarios involving network structure, initially infected nodes, action budgets, etc. It represents a combined intervention method based on structure-aware dynamic planning, providing a systematic learning and optimization pathway for implementing efficient and robust interventions in complex, time-varying public opinion networks.

Berger et al. [24] focused on comparing intervention methods and assessing their effects in false information governance. Through a large-scale randomized questionnaire survey experiment, they compared the short-term and immediate effects of fact-checking versus media literacy interventions. The study found that the effect of fact-checking was largely confined to the specific false information it directly corrected, whereas media literacy intervention more broadly helped users distinguish between true and false information, enhancing their ability to critically evaluate social media content, with effects persisting for about two

weeks. This research reveals the advantages of media literacy interventions in enhancing users' autonomous discrimination ability and improving intervention sustainability, and suggests media literacy as a cheap, scalable, easily implementable complementary means. This study exemplifies evidence-based comparison and selection of intervention strategies through experiments and effect evidence, providing empirical basis for optimizing and combining intervention methods in public opinion governance.

These studies, from different technical dimensions such as adaptive dynamic response, temporal assessment and delayed intervention, and structure-aware dynamic planning, promote the evolution of the public opinion control intervention stage from static, experience-driven approaches towards dynamic, algorithm-driven, effect-measurable, and adaptive learning intelligent regulation. They emphasize the quantitative modeling, real-time optimization, and systematic iteration of intervention strategies, providing multi-level methodological support and empirical foundation for constructing a scientific, precise, and actionable public opinion intervention system.

3.5. Fact-checking

In the fact-checking stage of intelligent public opinion governance, related research is dedicated to building automated systems capable of rapidly and accurately verifying the authenticity of social information and extracting supporting evidence, to curb the spread of false information and provide reliable evidence for subsequent governance decisions. In recent years, with the development of technologies such as Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and causal inference, fact-checking research has evolved from reliance on manual annotation towards intelligent, explainable, evidence-oriented approaches, enabling rapid response and efficient verification of real-time public opinion.

Manchanayaka et al. [25], from the perspective of causal inference, proposed using Convergent Cross Mapping (CCM) to identify coordinated attack behaviors on social media. This research treats user activity trajectories as time series and detects causal influence relationships between users through the CCM method, thereby revealing potential coordinated dissemination networks. Furthermore, they introduced topic modeling as a pre-clustering strategy, grouping users by discussion topics before conducting causal detection within groups, significantly improving computational efficiency and identification accuracy. This method not only effectively identifies coordinated dissemination behaviors but can also detect key influential nodes in the network, providing a causality-based detection means for identifying organized false information dissemination in public opinion governance.

Fionda [26] proposed a logic-based framework for analyzing fake news diffusion, formally modeling the propagation patterns of true and false news through News Diffusion Temporal Logic. This research characterizes news dissemination on social media as diffusion graphs with timestamps, extracts temporal dynamic features such as initial diffusion speed, lifecycle, and propagation concentration of news, and automatically generates explainable temporal logic rules to describe differences in propagation behaviors between false and true news. This rule-based logical framework not only improves early detection capability for fake news but also provides high explainability, offering a new methodological foundation for fact-checking based on information dissemination patterns in public opinion governance.

Singhal et al. [27] focused on enhancing the performance of LLMs on fact-checking tasks using RAG and few-shot in-context learning. They designed a pipeline consisting of document retrieval, evidence extraction, and veracity classification, dynamically retrieving relevant documents from a knowledge base, generating evidence question-answer pairs based on those documents, and finally having the LLM perform veracity classification based on the evidence. This method provides evidence-backed classification results, demonstrating the effectiveness of RAG and LLMs in automated fact-checking, particularly suitable for real-world scenarios lacking large-scale labeled data.

Tan et al. [28], addressing the need for causal reasoning in accident investigation reports, proposed a method for constructing causal knowledge graphs using LLMs and selecting nodes using Graph Neural Networks (GNNs). By extracting and inferring causal relationships from LLMs and organizing them in the form of knowledge graphs, combined with graph retrieval to enhance LLM reasoning capability, they significantly improved the accuracy of causal reasoning-based fact-checking, providing structured reasoning support for tracing misinformation in complex public opinion events.

Rolinger and Liu [29] constructed an iterative fact-checking framework based on Graph-of-Thoughts (GoT), achieving dynamic expansion of reasoning paths through multiple rounds of question generation and evidence retrieval. Through effective iteration strategies and efficient implementation under the constraints of open models, this method demonstrates multi-path, multi-evidence fusion capability in fact-checking, suitable for complex information verification scenarios like large-scale social media.

Rosenbaum et al. [30] proposed a hybrid fact-checking system integrating knowledge graphs, LLMs, and real-time search agents. Through a knowledge-graph-first retrieval strategy ensuring high precision and explainability, and invoking real-time search for supplementation when knowledge graph coverage is insufficient. This system achieved high F1 scores on benchmarks like FEVER and successfully mined additional evidence in information-insufficient categories, demonstrating the coverage and explanatory advantages of combining structured knowledge sources with open-domain retrieval.

Hu et al. [31], focusing on early fake news detection, proposed a Multimodal Prompt Learning (MPL) framework based on the multimodal pre-trained model CLIP. Through learnable prompt vectors and a multimodal feature fusion module, this model efficiently integrates image and text information and performs classification even with few labeled samples, achieving high recall and precision particularly in veracity alert tasks, providing a lightweight, rapid multimodal fact-checking tool for early public opinion intervention.

These studies, from different technical dimensions, jointly promote the evolution of fact-checking from content-based single verification towards a multidimensional, intelligent direction integrating propagation dynamics, evidence retrieval, causal reasoning, multimodal analysis, and iterative reasoning. They provide diversified technical pathways for constructing a scientific, efficient, and explainable fact-checking system within intelligent public opinion governance.

4. Summary

This paper systematically reviews the research progress of

intelligent public opinion governance from theoretical, methodological, and practical perspectives, centered on a core issue within the secondary discipline of Security and Protection Engineering. Grounded in public opinion governance theory, the paper analyzes the generation mechanisms, evolution characteristics of public opinion risks in the social media environment, and their challenges to public safety, clarifying the theoretical positioning of intelligent public opinion governance in risk warning and emergency response. It proposes and elaborates a fundamental framework for intelligent public opinion governance, constructing a methodological system covering five stages: "pre-event warning—early identification—situation prediction—control intervention—fact-checking," providing a structured, actionable research pathway for public opinion governance. Focusing on the disciplinary hotspot of intelligent public opinion governance on social media, the paper systematically reviews representative achievements and technical methods in the five key stages from recent top-tier domestic and international research. Through this systematic review of frontier research, the paper demonstrates the research trend in this field evolving from experience-orientation towards data-driven, model-optimized, and intelligent decision-making. It provides systematic theoretical references, methodological reference, and technical insights for public opinion governance research under the discipline of Security and Protection Engineering, contributing to promoting the scientific, precise, and intelligent development of intelligent public opinion governance systems in public safety practice.

References

[1] Suh J, Jahanparast E, Moon S, et al. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions[C]//80th Annual AAPOR Conference. AAPOR, 2025.

[2] Wang J, Yin Y, Wei L. Modeling public opinion dynamics in social networks using a GAN-SEIR framework[J]. Social Network Analysis and Mining, 2025, 15(1): 40.

[3] Huo Q, Zhang L, Zheng Q. Prediction of Public Opinion Event Types Combining Retrieval-Augmented Generation and Large Language Models[C]//Proceedings of the 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy. 2025: 392-398.

[4] Li Y, Garg K, Caragea C. A new direction in stance detection: Target-stance extraction in the wild[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 10071-10085.

[5] Gatto J, Sharif O, Preum S. Chain-of-thought embeddings for stance detection on social media[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 4154-4161.

[6] Li A, Liang B, Zhao J, et al. Stance detection on social media with background knowledge[C]//Proceedings of the 2023 conference on empirical methods in natural language processing. 2023: 15703-15717.

[7] Zhang R, Tian Y, Wei P, et al. An LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance identification[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 12253-12266.

[8] El-Mefleh M A, Alqaisi F. Building Social Early Warning System (SEWS): Predicting Social Unrest Through Economic Early Warnings[J]. Journal of Cultural Analysis and Social Change, 2025: 2326-2336.

[9] Kim R M, Veselovsky V, Anderson A. Capturing dynamics in online public discourse: A case study of universal basic income discussions on reddit[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2025, 19: 1021-1037.

[10] Parekh T, Mac A, Yu J, et al. Event Detection from Social Media for Epidemic Prediction[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 5758-5783.

[11] Shang L, Zhang Y, Yue Z, et al. A domain adaptive graph learning framework to early detection of emergent healthcare misinformation on social media[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2024, 18: 1408-1421.

[12] Hu W, Wang Y, Jia Y, et al. A multi-modal prompt learning framework for early detection of fake news[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2024, 18: 651-662.

[13] Martín-Corral D, García-Herranz M, Cebrian M, et al. Social media sensors as early signals of influenza outbreaks at scale[J]. EPJ Data Science, 2024, 13(1): 43.

[14] Feng Z, Yang Y, Huang X, et al. Efficient sphere-effect based information diffusion prediction on large-scale social networks[C]// Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025: 615-625.

[15] Wang R, Lu T, Zhang P, et al. Data-Driven Agent-Based Model for Public Opinion Propagation Simulation in Cyberbullying[J]. Big Data Mining and Analytics, 2025, 8(4): 794-819.

[16] Donkers T, Ziegler J. Understanding Online Polarization Through Human-Agent Interaction in a Synthetic LLM-Based Social Network[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2025, 19: 457-478.

[17] Cisneros-Velarde P. Biases in Opinion Dynamics in Multi-Agent Systems of Large Language Models: A Case Study on Funding Allocation[C]//Findings of the Association for Computational Linguistics: NAACL 2025. 2025: 1889-1916.

[18] Zhong T, Zhang J, Cheng Z, et al. Information diffusion prediction via cascade-retrieved in-context learning[C]// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2472-2476.

[19] Jin R, Liu X, Murata T. Predicting popularity trend in social media networks with multi-layer temporal graph neural networks[J]. Complex & Intelligent Systems, 2024, 10(4): 4713-4729.

[20] Wang G, Zhang R, Zhang Z. Opinion Maximization in Social Networks by Modifying Internal Opinions[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems.

[21] Chu Y. Dynamic response and disposal strategies for public opinion crises driven by reinforcement learning[J]. Discover Artificial Intelligence, 2025.

[22] Ghosh S, Mitra P, Nakov P. Clock against chaos: dynamic assessment and temporal intervention in reducing misinformation propagation[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2024, 18: 462-473.

[23] Muppasani B, Nag P, Narayanan V, et al. Towards effective planning strategies for dynamic opinion networks[J]. Advances in Neural Information Processing Systems, 2024, 37: 137046-137104.

[24] Berger L M, Kerkhof A, Mindl F, et al. Debunking “fake news” on social media: Immediate and short-term effects of fact-

checking and media literacy interventions[J]. Journal of Public Economics, 2025, 245: 105345.

[25] Manchanayaka I, Zaidi Z R, Karunasekera S, et al. Using causality to infer coordinated attacks in social media[C]// Proceedings of the International AAAI Conference on Web and Social Media. 2025, 19: 1176-1189.

[26] Fionda V. Logic-based analysis of fake news diffusion on social media[J]. Social Network Analysis and Mining, 2025, 15(1): 59.

[27] Singal R, Patwa P, Patwa P, et al. Evidence-backed fact checking using RAG and few-shot in-context learning with LLMs[C]//Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). 2024: 91-98.

[28] Tan F A, Desai J, Sengamedu S H. Enhancing fact verification with causal knowledge graphs and transformer-based retrieval for deductive reasoning[C]//Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). 2024: 151-169.

[29] Rolinger S, Liu J. Graph-of-thoughts for fact-checking with large language models[C]//Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER). 2025: 266-273.

[30] Rosenbaum R, Cavelius T, Strothe L, et al. Hybrid Fact-Checking that Integrates Knowledge Graphs, Large Language Models, and Search-Based Retrieval Agents Improves Interpretable Claim Verification[C]//Proceedings of the 9th Widening NLP Workshop. 2025: 106-115.

[31] Hu W, Wang Y, Jia Y, et al. A multi-modal prompt learning framework for early detection of fake news[C]Proceedings of the International AAAI Conference on Web and Social Media. 2024, 18 651-662.