UDC: 004.93 https://doi.org/10.15407/jai2025.04.108

O. Ostrovska[1], V. Lyashkevych[2]
[1,2] Ivan Franko National University of Lviv, Ukraine
50 Drahomanova str., Lviv, Lviv 79005
[1] oksana.ostrovska@lnu.edu.ua
[2] vasyl.liashkevych@lnu.edu.ua
[1]https://orcid.org/0009-0006-3376-8448
[2]https://orcid.org/0000-0003-2810-6061

AUTOMATED ONLINE DETECTION OF HARMFUL AND DANGEROUS CONTENT IN SOCIAL NETWORKS: A SYSTEMATIC REVIEW

Abstract. The proliferation of social networks has transformed global communication, yet it has concurrently facilitated the rapid and widespread dissemination of harmful content. This content, encompassing disinformation, hate speech, extremism, and cyberbullying, poses significant threats to individual well-being, social cohesion, and democratic processes. The sheer volume and velocity of user-generated data render manual moderation untenable, necessitating the development of sophisticated automated detection systems - therefore, the social networks are one of the best environments for swindlers, cheaters, and other inadequate people to spread off harmful and dangerous content. This review systematically analyzes scientific literature to map the landscape of existing solutions and identify critical gaps for future research.

A thorough literature review was conducted following formal protocol. Key scholarly databases such as IEEE Xplore, ACM Digital Library, Scopus, and arXiv were searched using a comprehensive set of keywords related to malicious content, social media, and detection methodologies. The review focused on peer-reviewed articles, conference proceedings, and scholarly books published in the last five years, supplemented by foundational works.

The analysis reveals a clear evolution in detection methodologies, from traditional machine learning (ML) models reliant on manual feature engineering to advanced deep learning architectures. A taxonomy of harmful content is established, clarifying the distinctions between phenomena such as disinformation, extremism, and cyberbullying. The review critically examines the dominant detection paradigms: content-based methods using Natural Language Processing (NLP), structure-based methods leveraging Graph Neural Networks (GNNs) to analyze propagation patterns, and emerging hybrid and multi-modal approaches. Despite significant progress, all current methods face fundamental limitations, including a critical lack of contextual understanding, susceptibility to algorithmic bias, vulnerability to adversarial attacks, and a pervasive lack of transparency and explainability.

Current methodologies for harmful content detection, while increasingly sophisticated, remain largely reactive and operate on static snapshots of data. They are insufficient for the early, robust, and context-aware identification of evolving threats. A significant research gap exists for a new generation of systems that deeply integrate dynamic content and network analysis. Future research should focus on developing proactive solutions founded on temporal graph learning and Complex Event Processing (CEP), with explainability integrated by design, to effectively model, detect, and mitigate harmful scenarios as they unfold in real-time.

Keywords: harmful content detection, social network analysis, natural language processing, graph neural networks, information diffusion, content moderation.

Introduction

The 21st century has been defined by the rise of the digital social sphere. Platforms such as Facebook, X (Twitter), TikTok, and YouTube have reshaped human interaction, creating a globally interconnected network where information, ideas, and culture are exchanged with unprecedented speed and scale. These platforms are not merely communication channels; they are complex, dynamic information environments that actively shape public discourse, influence social behavior, and constitute a new form of social reality [11, 13]. The ability for any user to generate and disseminate content to a potential global audience has been heralded as a democratization of communication. However, this same capability presents a profound, double-edged sword. The very openness that facilitates connection and expression has also created a fertile ground for the proliferation of harmful content. The spectrum of this harm is broad, ranging from coordinated disinformation campaigns that erode trust in democratic institutions to hate speech that incites violence, and from cyberbullying that inflicts severe psychological trauma to the normalization of self-harm among vulnerable populations [17].

The volume and velocity of this user-generated content are staggering; with billions of users active daily, the torrent of new posts, images, and videos makes manual oversight

and moderation a logistical impossibility. This "question of scale" has driven an urgent need for automated systems capable of identifying and mitigating harmful content before it can cause widespread damage [12, 15]. The grand challenge, however, extends far beyond simple classification. Identifying harmful content is not a straightforward technical problem, but one that lies at the intersection of computational linguistics, network science, and human psychology. Harm is often context-dependent, linguistically nuanced, and culturally specific. Malicious actors continuously devise new methods to evade detection, while the very algorithms designed by platforms to maximize user engagement can inadvertently amplify the most sensationalist, polarizing, and ultimately harmful content [6]. This creates a fundamental tension where the architectural and economic incentives of a platform can be at odds with the safety of its users. Consequently, developing effective solutions requires a holistic understanding that addresses the content itself, the structure of the network through which it spreads, and the dynamic behaviors of the users who create and consume it.

This systematic review aims to provide a comprehensive and critical analysis of the scientific literature on harmful content identification in online social networks. The primary objectives of this paper are fourfold:

– To systematically survey and synthesize the peer-reviewed literature, establishing a foundational understanding of the problem domain.

– To develop a clear, scientifically grounded taxonomy of harmful content types and their defining characteristics, moving beyond colloquialisms to establish precise terminology.

– To critically analyze and compare the primary methodologies employed for detection, including content-based, structure-based, and hybrid approaches, while also examining the theoretical models of information diffusion that explain their spread.

– To identify the critical limitations and unresolved challenges of current solutions, thereby delineating a precise and defensible research gap that can guide the development of a novel, more effective generation of detection systems for a doctoral-level contribution.

By achieving these aims, this review serves as both a comprehensive survey for the academic community and a foundational analysis for a new research trajectory focused on building more robust, context-aware, and proactive solutions to one of the most pressing technological and societal challenges of our time.

Implementation

Online social networks are far more than passive conduits for messages. They are intricately designed, algorithmically curated ecosystems that actively shape user experience and information flow. Literature increasingly frames these platforms as "information environments," defined by Floridi as systems "constituted by all informational processes, services, and entities, thus including informational agents as well as their properties, interactions, and mutual relations" [11]. In this context, users are "informational agents" whose behavior is influenced by, and in turn shapes, the environment.

The structure of these networks is a critical determinant of their function. Most large-scale social networks exhibit properties like scale-free degree distributions (a few highly connected hubs and many sparsely connected nodes) and small-world phenomena (short path lengths between any two users), which facilitate rapid information propagation. However, the most significant architectural feature is the use of algorithmic content curation. Given the overwhelming volume of content, platforms do not present information chronologically. Instead, they employ sophisticated algorithms to filter and rank content based on predicted relevance and potential for user engagement. This algorithmic mediation, while intended to improve user experience, is a primary factor in the amplification of harmful content. These systems are typically optimized for business metrics such as clicks, shares, comments, and time-on-site. Unfortunately, content that is shocking, polarizing, outrageous, or emotionally charged - hallmarks of much disinformation and extremist rhetoric - is often highly engaging [6]. This creates a perverse incentive structure: the very algorithms designed to promote content can inadvertently

favor and accelerate the spread of harm. The platform's architecture, therefore, is not a neutral backdrop but an active participant in the problem, creating feedback loops where harmful content is rewarded with greater visibility, leading to a wider audience and further engagement. Any robust solution to harmful content must contend with this fundamental, system-level dynamic.

A precise taxonomy of harmful content is essential for detection. Based on the literature, it can be classified by intent, veracity, and target.

Information disorders concern the truthfulness and intent of information. Misinformation is "misleading information that is created and spread, regardless of whether there is an intent to deceive". Disinformation is a subset of misinformation created and spread with the explicit intent to deceive. Fake News is a form of disinformation characterized as "deliberate and verifiable false news" that often mimics legitimate news sources [4, 17]. Malinformation is genuine information shared out of context with the intent to cause harm [11].

Ideologically motivated harm involves content promoting extreme ideologies. Online Extremism includes speech and actions that marginalize or denigrate groups and can be a precursor to terrorism. Extremist groups often use coded language to evade detection. The architecture of social media is conducive to extremism, as anonymity can embolden radical views and algorithmic recommendations can create "echo chambers" that lead to polarization and radicalization [6].

Interpersonal harm focuses on harm directed between individuals or groups. Cyberbullying is distinguished by repetition and intent, defined as a repeated activity via electronic means intended to cause psychological harm. Its forms include harassment, denigration, impersonation, outing, and exclusion. It is considered uniquely pernicious due to potential anonymity, pervasiveness, large potential audience, and lack of supervision [25]. Hate Speech is language that attacks or demeans a person or group based on protected attributes like race, religion, or gender [12]. A key challenge is the ambiguity between hate speech and merely offensive language. Incitement to Self-Harm includes content that normalizes, glorifies, or provides instruction for self-injury, which can create a contagion effect. Dangerous Scenarios and Illicit Content is a broad category including viral "challenges," scams, and the non-consensual sharing of explicit or violent content. To provide a clear reference, these categories are summarized in Table 1.

On a societal level, the impact is equally corrosive. The widespread dissemination of disinformation and "fake news" erodes public trust in core institutions such as the media, science, and government [11, 17]. This can have dire consequences, from influencing election outcomes to undermining public health initiatives. Extremist content and polarizing rhetoric fuel social fragmentation and have been directly linked to an increase in hate crimes and real-world political violence [6]. The creation of online echo chambers, where individuals are shielded from opposing viewpoints, accelerates this polarization, making constructive public discourse increasingly difficult [6]. The cumulative effect is a degradation of the shared information environment upon which a functioning democracy depends.

To effectively stop the spread of harmful content, it is first necessary to understand the mechanisms by which it propagates. Information diffusion modeling is the field of study dedicated to this task. It seeks to create mathematical and computational models that describe how information, ideas, and behaviors spread through a network. The evolution of these models reveals a growing appreciation for the complexity of human social dynamics.

The approaches to modeling information diffusion have evolved significantly, moving from broad, population-level analogies to more granular, behaviorally nuanced frameworks.

A pivotal finding in the field, and the primary justification for structure-based detection methods, is that harmful content does not spread in the same way as benign content. Empirical studies on large-scale social media data have consistently shown that false news, in particular, diffuses "significantly farther, faster, deeper, and more broadly than the truth in all categories of information" [17]. This difference in propagation signature provides a powerful,

non-content-based signal for detection. While true news tends to spread more slowly, often through a few large broadcast events, false news cascades are typically characterized by a greater number of smaller diffusion chains initiated by many different users. This suggests that fake news is more "viral" in the sense that it relies more on peer-to-peer sharing by less influential users. The structure of these cascades also holds clues. Some research indicates that the breadth of a cascade is a more reliable predictor of its ultimate size than its depth [14]. This implies that content that quickly captures a wide, diverse audience is more likely to become a massive cascade.

Table 1. Taxonomy of Harmful Content in Social Networks

Content Type	Definition	Key Characteristics	Examples
Misinformation	False or inaccurate information spread without malicious intent.	Veracity: False; Intent: Benign/Unknowing.	"A user sharing an old, inaccurate weather warning believing it is current."
Disinformation	False information is deliberately created and spread to deceive or cause harm.	Veracity: False; Intent: Malicious.	A state-sponsored campaign using bots to spread fabricated stories about a political opponent.
Fake News	A form of disinformation that mimics the format and style of legitimate news.	Veracity: False; Intent: Malicious; Format: Journalistic.	An article on a fake news website with a misleading headline claiming a celebrity has died.
Malinformation	Genuine information shared out of context to mislead or cause harm.	Veracity: True; Intent: Malicious.	Sharing a real photo of a protest from a different country to falsely represent a local event.
Extremism	Speech or actions promoting extreme ideologies, often targeting marginalized groups.	Ideological; Polarizing; Can be non-violent but is a precursor to violence; Often uses coded language.	A forum post advocating for racial segregation using veiled terminology.
Hate Speech	Abusive language targets individuals or groups based on protected characteristics.	Targeted; Demeaning; Based on identity (race, religion, gender, etc.).	A tweet containing racial slurs directed at an individual.
Cyberbullying	Repeated and intentional harm inflicted through electronic means.	Repetitive; Intentional; Power Imbalance; Psychological Harm.	A series of harassing comments and threats sent to a classmate's profile over several days.
Self-Harm Content	Content that promotes, glorifies, or provides instruction for self-injury.	Normalizing; Instructional; Contagious.	A live video demonstrating acts of self-cutting.
Dangerous Scenarios	Content encouraging risky behaviors, scams, or exposure to illicit material.	Risk-taking; Deceptive; Inappropriate.	A viral "challenge" involving misuse of household chemicals; a romance scam.

The existence of these distinct propagation patterns is fundamental. If all information spread identically, detection would have to rely solely on analyzing the content of the message. However, the fact that the structure of the interaction graph itself contains information about the veracity or nature of the content being spread opens up an entirely new dimension for detection. This insight forms the critical bridge between the theoretical study of information diffusion and the practical challenge of building detection systems, justifying the graph-based methodologies discussed in the following section.

The academic and industrial response to the challenge of harmful content has produced a diverse arsenal of detection methodologies. These approaches have evolved from simple, rule-based systems to highly complex deep learning architectures that attempt to model language, social structure, and user behavior simultaneously.

The earliest automated systems relied on traditional machine learning classifiers fed with

manually engineered features. This pipeline involves extracting specific, predefined characteristics from the text and using them to train models like Support Vector Machines (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forests. Common features include: Lexical Features, Syntactic Features, Statistical Features (n-grams, TF-IDF, Bag-of-Words), and Sentiment Features. Strengths of these models are their interpretability and lower computational cost. Their weaknesses are a reliance on hand-crafted features, which struggle to capture semantic nuance and are brittle in the face of evolving language.

Deep learning in NLP enabled models to learn features automatically from raw text. Convolutional Neural Networks (CNNs) are effective for detecting local patterns in text [4]. Recurrent Neural Networks (RNNs), like LSTM and GRU, are designed for sequential data, capturing long-range dependencies and context [4]; bidirectional versions are particularly powerful [4]. The strength of these deep learning models is their ability to outperform traditional ML by learning complex language representations without manual feature engineering. Their weaknesses include being "black boxes" with low interpretability, requiring large, labeled datasets, and being computationally expensive.

Transformers, such as BERT and RoBERTa, are the current state of the art in NLP [8, 16]. Their key innovation is the self-attention mechanism, which allows the model to weigh the influence of all words simultaneously for a deeper contextual understanding [1]. These models are pre-trained on massive text corpora and then fine-tuned for specific tasks [8, 16]. Their strength is achieving state-of-the-art accuracy on tasks like harmful content detection [3, 16]. Their weaknesses include being even more computationally demanding and data-hungry, and they are blind to extra-textual information like user credibility or network propagation patterns [3].

Structure-based methods shift focus from the content to the speaker and the propagation dynamics. Graph Neural Networks (GNNs) are a class of deep learning models for graph-structured data [16, 18]. A GNN learns a feature representation for each node by iteratively aggregating features from its neighbors, allowing it to learn from both node features and the graph's topology. GNNs are effective for fake news detection by modeling propagation graphs (e.g., with models like BiGCN) to classify them based on structural properties [7]. The strength of GNNs is their ability to model complex relational data and dynamics that are invisible to content-only NLP models [18]. Their weaknesses include potentially lower performance on tasks dependent on linguistic nuance and technical issues like over-smoothing in large graphs [3].

Recognizing the limitations of unimodal approaches, the research frontier has moved towards models that combine multiple sources of information.

These models represent the most promising direction for holistic detection, aiming to get the best of both worlds by combining NLP and GNN techniques. A typical hybrid architecture consists of a content branch (e.g., BERT) and a structure branch (a GNN). The representations from these two branches are then fused and fed into a final classifier [5, 10]. StopHC architecture is a prime example of this philosophy, proposing a harmful content detection module and a separate network immunization module to mitigate spread.

Harmful content is not limited to text. Hateful memes, deepfake videos, and graphic images are potent forms of harm. Multi-modal detection addresses this by analyzing information from different modalities (text, image, audio). These models use a specialized neural network for each modality, and then the feature representations are combined using a fusion strategy to make a final prediction [2]. The primary challenges in multi-modal detection are developing effective fusion mechanisms and gracefully handling missing modalities.

While the models above focus on classifying content, Complex Event Processing (CEP) offers a higher-level paradigm for building detection systems. CEP is a technology for analyzing high-volume streams of data (events) in real-time to identify meaningful patterns, or "complex events". In this context, a "simple event" can be a user post, a "like," or a share. A "complex event," which the system is designed to detect, could be a

composite pattern like: "A new account posts content with a high toxicity score that is rapidly shared by a cluster of accounts with low credibility, indicating the start of a coordinated disinformation campaign" [5]. CEP systems are inherently designed for real-time, low-latency processing. This shifts the paradigm from reactive, post-by-post classification to proactive, scenario-based monitoring and intervention. This approach directly addresses the critical need for the early detection and mitigation of spreading harm.

To transition from a general overview of methodologies to identifying concrete research opportunities, it is instructive to analyze the performance and limitations of specific architectures presented in the literature. This comparative analysis highlights the trade-offs inherent in current solutions and pinpoints the most promising gaps for novel improvements.

This detailed comparison reveals a clear pattern: while individual models excel in specific areas - Transformers in language, GNNs in structure, and TGNs in dynamics, they often operate with significant blind spots. The most sophisticated hybrid models attempt to combine these strengths, but their fusion is often shallow, and they still operate reactively on static data points. This analysis reinforces the conclusion that the most significant and impactful research gap lies in the development of a system that can proactively detect unfolding harmful scenarios by deeply and dynamically integrating content, structure, and temporality in a unified, explainable framework.

Despite the rapid evolution of detection technologies, the field is constrained by a set of persistent and fundamental challenges. These limitations are not merely technical hurdles to be overcome with more data or bigger models; they point to deep-seated complexities in language, human behavior, and social systems. Synthesizing these challenges is crucial for identifying the most impactful directions for future research.

Across the literature, several recurring themes highlight the fundamental gap between automated capabilities and the requirements of effective, fair moderation.

The most frequently cited limitation of automated systems is their profound difficulty in understanding context [1]. Language is not a static code; its meaning is fluid and heavily dependent on cultural norms, the relationship between speakers, and the specific situation. An algorithm may easily flag a reclaimed slur used within a marginalized community as hate speech, while failing to detect a sophisticated, sarcastic, or ironic jab that carries a harmful subtext. What one culture considers offensive, another may see as benign. Because harm is an inherently subjective and socially constructed concept, creating a universal, objective classifier is a fundamentally ill-posed problem. This implies that even a technically perfect system will inevitably make errors that appear nonsensical to a human observer, leading to both false positives (wrongful censorship) and false negatives (missed harm) [13].

A critical ethical and technical challenge is that machine learning models are susceptible to inheriting and amplifying biases present in their training data. If a dataset contains historical biases where, for example, the speech of a particular demographic group is disproportionately labeled as "toxic," a model trained on this data will learn to replicate that bias. This can lead to discriminatory outcomes where automated moderation systems unfairly target and censor already marginalized communities, exacerbating the very harms they are meant to prevent [9]. Auditing and mitigating this algorithmic bias is a major area of ongoing research, but it remains a significant obstacle to the fair deployment of these systems.

Harmful content is not a static target. Malicious actors are engaged in a continuous, adaptive effort to evade detection. They employ tactics like using lexical variations (e.g., "h8" for "hate"), coded language ("dog whistles"), manipulated images, and sophisticated deepfakes to bypass automated filters [4]. As soon as a detection system learns to identify one pattern, adversaries develop a new one. This creates a perpetual cat-and-mouse game, rendering static models trained on historical data quickly obsolete. Effective systems must therefore be dynamic and capable of continuous learning and adaptation.

Table 2. AI models

Methodology	Core Principle	Primary Strengths	Key Limitations
Traditional ML	Relies on manually engineered features (lexical, syntactic, etc.) to train classifiers.	Interpretable and computationally less expensive.	Struggles to capture semantic nuance; brittle against evolving language and adversarial attacks.
Deep Learning	Automatically learns feature representations from raw text. RNNs capture sequential context, while CNNs identify local patterns.	Achieves superior performance over traditional ML by learning complex features without manual engineering.	Often a "black box" with low interpretability; requires large, labeled datasets; computationally expensive.
Transformers	Uses a self-attention mechanism to weigh all words simultaneously, enabling a deep contextual understanding of language.	Achieves state-of-the-art accuracy on many NLP tasks, including harmful content detection.	Highly demanding in terms of data and computation; blind to extra-textual information like network structure or user credibility.
Graph Neural Networks	Models' information as a graph, learning from both node features (e.g., user profiles) and the graph's structure (i.e., propagation patterns).	Can detect harmful content based on propagation dynamics (how it spreads), which is invisible to content-only models.	May underperform on tasks that depend solely on linguistic nuance; can suffer from technical issues like over-smoothing in large graphs.
Hybrid & Multi-modal	Fuses information from multiple sources. This can involve combining content (NLP/BERT) and structure (GNN) or different data types (e.g., text, images).	Represents the most holistic approach by leveraging the strengths of different models to create a more robust detection system.	Developing effective fusion mechanisms is challenging; models often remain reactive and can struggle with missing data modalities.

The sheer scale of social media presents a dual challenge. The volume of content necessitates automation, yet the most powerful and nuanced models (like large Transformers or complex GNNs) are often too computationally expensive to deploy for real-time analysis of every single piece of content [15]. There is a constant trade-off between model complexity (and thus potential accuracy) and deployment feasibility. This challenge is compounded by the need for real-time moderation to prevent harmful content from going viral before it can be addressed [13].

A direct consequence of the increasing complexity of deep learning models is their lack of transparency. High-performance models like Transformers and GNNs are often "black boxes," meaning their internal decision-making processes are opaque even to their creators [1]. It is possible to see the input and the output, but not the reasoning that connects them. This opacity is a critical barrier to trust and accountability in a high-stakes domain like content moderation, which directly impacts fundamental rights like freedom of expression. Without understanding why, a model flagged a piece of content, it is impossible for a user to mount a meaningful appeal, for a platform to audit its systems for bias, or for regulators to ensure compliance with legal standards.

Explainable AI (XAI) is an emerging subfield of AI dedicated to addressing this problem. XAI seeks to develop techniques that can make model decisions interpretable to humans. These techniques range from post-hoc methods that try to explain a black box model's decision (e.g., by highlighting the input features that were most influential) to creating models that are inherently interpretable by design [10]. The application of XAI to content moderation is crucial for building systems that are not only accurate but also fair, accountable, and trustworthy.

This trade-off between predictive power and transparency underscores the importance of carefully selecting the right architecture for the task. The campaigns that undermine democratic institutions, hate speech that incites real-world violence, and cyberbullying that inflicts severe psychological trauma - poses a direct threat to individual well-being and social cohesion. The staggering volume and velocity of this user-generated data make manual moderation a logistical impossibility, creating

an urgent and non-negotiable need for effective automated systems.

Conclusion

The rise of online social networks has fundamentally reshaped human communication, creating an interconnected global sphere for the exchange of ideas and culture. However, this digital revolution carries a profound downside: the same open architecture that fosters connection also serves as a fertile ground for the rampant spread of harmful content. This content - a wide spectrum including coordinated disinformation campaigns that undermine democratic institutions, hate speech that incites real-world violence, and cyberbullying that inflicts severe psychological trauma - poses a direct threat to individual well-being and social cohesion. The staggering volume and velocity of this user-generated data make manual moderation a logistical impossibility, creating an urgent and non-negotiable need for effective automated systems.

This review has charted the evolution of automated detection methodologies, tracing a clear trajectory from traditional machine learning models, which depend on brittle, hand-crafted features, to the sophisticated deep learning architectures that define the current state of the art. The analysis critically examined the dominant paradigms: content-based methods using Natural Language Processing (NLP) models like BERT to understand text; structure-based methods using Graph Neural Networks (GNNs) to analyze the unique patterns by which harmful content propagates; and emerging hybrid and multi-modal approaches that attempt to create a more holistic view by combining these techniques. A scientifically grounded taxonomy was established, clarifying the distinctions between phenomena like misinformation, disinformation, hate speech, and extremism, each presenting unique challenges for detection.

However, a comprehensive survey of these advanced techniques reveals a sobering reality. Despite their increasing sophistication, all current methods suffer from fundamental limitations that prevent them from being a complete solution. They remain overwhelmingly reactive, designed to classify content after it has been posted, operating on static snapshots of data. More critically, they are constrained by a profound lack of contextual understanding, an inherent vulnerability to adversarial evasion, a susceptibility to amplifying societal biases, and a dangerous lack of transparency that erodes trust and accountability. These are not minor flaws but deep-seated challenges that mark the boundary of current capabilities.

Therefore, the most significant and defensible research gap lies in shifting the entire paradigm from reactive classification to proactive, scenario-based detection. The future of harmful content moderation does not lie in building a slightly more accurate classifier, but in developing a new generation of systems capable of identifying and mitigating threats as they unfold in real-time. This requires a deep integration of dynamic content and network analysis, moving beyond static data points to model the temporal evolution of online interactions.

The path forward points directly toward solutions founded on temporal graph learning and Complex Event Processing (CEP). Such a system would not just analyze a single post but would detect meaningful patterns from high-volume event streams - identifying, for instance, the complex event of a new account posting toxic content that is then rapidly shared by a cluster of low-credibility users, flagging a likely disinformation campaign at its inception. Furthermore, this new architecture must have explainability integrated by design, not as an afterthought. This is essential for building systems that are not only effective but also fair, auditable, and worthy of the public's trust. By pursuing this research trajectory, we can move beyond the current cat-and-mouse game and begin to build a digital information environment that is safer, healthier, and more resilient.

References

1. Roberts, S. T. (2019). Behind the screen: Content moderation in the shadows of social media. Yale University Press.

2. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks

and learning systems, 32(1), 4-24.

3. Al-Sarem, M., et al. (2024). Comparative Analysis of Graph Neural Networks and Transformers for Robust Fake News Detection: A Verification and Reimplementation Study. Electronics, 13(23), 4784.

4. Mathew, B., et al. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867-14875.

5. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.

6. Awan, I., Sutch, H., & Carter, B. (2019). Extremism Online: An analysis of the role of social media in the growth of far-right extremism. Centre for Analysis of the Radical Right.

7. Bian, T., et al. (2020). Rumor detection on social media with bidirectional graph convolutional networks. Proceedings of the AAAI conference on artificial intelligence, 34(01), 549-556.

8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

9. Dixon, L., et al. (2018). Measuring and mitigating unintended bias in text classification. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.

10. Truică, C. O., Constantinescu, A. T., & Apostol, E. S. (2024). StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms. arXiv preprint arXiv:2411.06138.

11. Treen, K. M., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. Wiley Interdisciplinary Reviews: Climate Change, 11(5).

12. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1-30.

13. Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

14. Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. Management Science, 62(1), 180-196.

15. Gorwa, R. (2019). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 6(1).

16. Kiela, D., et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems, 33, 2611-2624.

17. Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

18. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151

The article has been sent to the editors 21.11.25.
After processing 10.12.25.
Submitted for printing 30.12.25.