

Graph Neural Network-Driven Social Bot Detection

S.Guru Gowri¹, V.Madhava²

¹PG Scholar, Dept.of CSE, Sri Venkateswara Institute of Science and Technology, Kadapa.

²Associate Professor, Dept.of CSE, Sri Venkateswara Institute of Science and Technology, Kadapa.

Email: gowrisure2001@gmail.com, cemadhava@gmail.com

Abstract –

The rapid expansion of online social networks has fundamentally transformed how information is produced, disseminated, and consumed. While these platforms enable open communication and global connectivity, they have also become fertile ground for malicious automation in the form of social bots. Social bots are automated or semi-automated accounts designed to imitate human behavior, often deployed to manipulate public opinion, spread misinformation, amplify extremist narratives, or distort online discourse through coordinated activities. Traditional detection techniques have struggled to keep pace with the increasing sophistication of these bots, particularly as modern bot systems leverage advanced natural language generation and complex interaction strategies that closely resemble genuine human users. In this context, this research introduces LGB, a novel social bot detection framework that integrates Language Models (LMs) with Graph Neural Networks (GNNs) to jointly analyze linguistic content and relational network structures. The proposed approach addresses the inherent limitations of content-only or structure-only detection systems by modeling both semantic intent and coordinated behavior patterns. By capturing deep contextual language representations alongside graph-based dependencies among users, posts, and interactions, LGB aims to deliver robust, scalable, and adaptive bot detection performance in dynamic social media environments.

Keywords – Neural Network, GNN, Language Models, Bot detection, Social Network.

1. INTRODUCTION

Social media platforms have become central to political discourse, public health communication, marketing strategies, and social interaction. Alongside their benefits, these platforms have witnessed a dramatic rise in automated entities known as social bots. Early generations of bots were relatively simple, often identified through abnormal posting frequency or rule-based behavioral patterns. However, recent advances in artificial intelligence, particularly in natural language processing and automated decision-making, have enabled bots to generate highly fluent text, engage in conversations, and mimic nuanced human behavior. This evolution has rendered many traditional bot detection systems ineffective, as they rely on shallow statistical features or manually crafted rules that fail under adversarial adaptation. Existing detection systems frequently focus on isolated perspectives, such as linguistic cues or user activity metrics. While such approaches can identify obvious automation, they struggle with coordinated bot networks where individual accounts may appear human-like in isolation but reveal malicious intent through collective behavior. Furthermore, the dynamic and adversarial nature of social platforms demands detection models that can generalize across topics, languages, and evolving bot strategies. These challenges highlight the need for integrated approaches capable of reasoning about both what is being said and how information propagates through social structures. This research responds to that need by proposing a hybrid architecture that combines language-level semantic

understanding with graph-based relational learning, offering a comprehensive framework for modern social bot detection.

2. EXISTING SYSTEM

Support Vector Machines (SVMs) have historically been among the most widely adopted machine learning algorithms for social bot detection due to their strong theoretical foundation and effectiveness in high-dimensional feature spaces. In traditional systems, SVMs are trained using handcrafted features derived from user profiles, posting behavior, temporal activity patterns, and basic linguistic statistics such as word frequency or sentiment polarity. These features are manually selected based on domain expertise and fed into the SVM classifier to distinguish between human and bot accounts. While SVM-based systems demonstrated promising results during early stages of social bot research, their effectiveness has diminished as bot behaviors have become increasingly sophisticated. The reliance on static, manually engineered features limits the model's ability to capture deeper semantic meaning or adapt to new bot strategies. Moreover, SVMs treat each instance independently, failing to exploit relational information among users that is crucial for identifying coordinated bot networks. As a result, SVM-based approaches often perform well on controlled datasets but struggle in real-world, dynamic social media environments where bots actively adapt to detection mechanisms.

Disadvantages of Existing System

One of the primary limitations of SVM-based bot detection systems lies in their dependence on handcrafted features,

which are costly to design, dataset-specific, and vulnerable to adversarial manipulation. As bots adopt more human-like posting patterns and leverage advanced language generation techniques, shallow lexical or behavioral features lose their discriminative power. Additionally, SVMs lack inherent mechanisms for modeling sequential context or long-range dependencies in text, leading to poor performance when dealing with nuanced or context-dependent language. Another critical drawback is the inability of SVMs to model relational and collective behavior. Social bots often operate in coordinated groups, amplifying specific narratives through synchronized actions. SVM based systems, which classify accounts independently, fail to capture such coordination, resulting in high false-negative rates. Scalability also poses a challenge, as training SVMs on massive, continuously evolving social media datasets can be computationally expensive and inflexible. These limitations collectively motivate the transition toward deep learning-based architectures capable of learning rich representations from both content and structure.

3. PROPOSED SYSTEM

The proposed LGB framework integrates Language Models (LMs) with Graph Neural Networks (GNNs) to overcome the fundamental shortcomings of traditional bot detection systems. Language models are employed to extract deep semantic representations from textual content, capturing contextual meaning, syntactic structure, and subtle linguistic cues that distinguish human communication from automated generation. These representations enable the system to analyze not just surface-level word usage but also intent, coherence, and discourse patterns. Simultaneously, graph neural networks are utilized to model the relational structure of social media platforms. Users, posts, and interactions are represented as nodes and edges within a graph, allowing the model to learn patterns of coordination, influence, and information diffusion. By propagating information across the graph, GNNs can identify clusters of accounts that exhibit synchronized behavior indicative of bot networks. The fusion of language-level and graph-level representations enables LGB to perform holistic reasoning, where linguistic signals are interpreted within the context of social interactions. This joint learning paradigm significantly enhances robustness against adversarial evasion and improves generalization across different bot behaviors and domains.

Advantages of proposed system

The LGB framework offers several key advantages over traditional bot detection approaches. By leveraging deep language models, the system can adapt to evolving linguistic patterns without manual feature engineering, making it resilient to advances in automated text generation. The incorporation of graph neural networks allows for explicit modeling of coordinated behavior, enabling the detection of

bot networks that would otherwise evade content-only classifiers. Furthermore, the proposed system supports scalability and adaptability, as both language models and GNNs can be fine-tuned or incrementally updated as new data becomes available. The unified architecture also facilitates explainability, as linguistic and structural signals can be analyzed jointly to understand detection decisions. Overall, LGB represents a comprehensive, future-ready solution for social bot detection, addressing both the semantic and relational dimensions of malicious automation in online social networks.

Non - Functional Requirements

Non - Functional Requirement (NFR), frame a product foundation's feature credits. They assess the product framework as per non-functional standards that are fundamental for its prosperity, like success, such as portability, security, responsiveness, and usability. An illustration of a nonfunctional need may be "the manner by which rapidly could I at any point load the site?" Lacking satisfaction of non-practical rules might prompt frameworks that can't meet client prerequisites. Nonpractical prerequisites give you the capacity to put limits or cutoff points on how the framework is planned across various cycles of the lithe build-up. For example, provided that skilled are in excess of 10,000 things marked in without an importance's delay, the scene should stack in three seconds. Similarly, however meaningful as a proficient necessity's rendering can be individual for non-utilitarian necessities.

4. METHODOLOGY

The increasing reliance on online social media platforms for communication, information dissemination, and public engagement has fundamentally reshaped digital interaction. However, this transformation has also enabled the rapid proliferation of automated social bots capable of imitating human behavior at scale. These bots are often deployed to manipulate public opinion, amplify misinformation, distort online trends, and undermine the authenticity of digital discourse. Traditional bot detection systems, which primarily rely on rule based mechanisms or shallow machine learning models, are no longer sufficient to counter these evolving threats. The proposed LGB framework, which integrates advanced language models with graph neural networks, is designed to address the multidimensional nature of social bot behavior by jointly analyzing linguistic patterns and network-level interactions. The system analysis phase establishes the conceptual foundation of this approach by examining the operational challenges, data characteristics, and functional requirements necessary for effective bot detection. From a system analysis perspective, social media environments generate massive volumes of heterogeneous data that include user-generated textual content, temporal activity logs, metadata attributes, and complex relational structures formed

through interactions such as follows, mentions, replies, and reposts. This data is inherently noisy, highly dynamic, and structurally complex. Human users exhibit diverse linguistic expressions, irregular posting behaviors, and organically evolving social relationships, while bots increasingly attempt to replicate these characteristics using automated scripts and pretrained text generation models. As a result, surface-level behavioral indicators and handcrafted features fail to capture the subtle distinctions between genuine users and sophisticated bots. The LGB system analysis identifies this challenge and emphasizes the necessity of a hybrid analytical approach capable of modeling both semantic content and relational dependencies in a unified learning framework. The limitations of existing detection systems further motivate the proposed approach. Content-based detection methods focus on textual cues such as keyword frequency, sentiment polarity, and syntactic patterns, but these features are easily manipulated by bots employing modern language generation techniques. Network-based methods analyze structural properties such as connectivity patterns, centrality measures, and interaction density, yet coordinated bot networks can artificially construct realistic social graphs to evade detection. The system analysis concludes that isolating content or structure leads to incomplete representations of user behavior. In contrast, integrating language models with graph neural networks enables the system to detect inconsistencies between what users say and how they interact within the social network, providing a more robust and resilient detection mechanism. The system architecture of the LGB framework is designed to operationalize this hybrid analytical vision through a modular, scalable, and extensible design. At a high level, the architecture consists of interconnected layers responsible for data ingestion, representation learning, feature fusion, classification, and decision output. Raw social media data enters the system through a data acquisition interface that aggregates textual posts, user metadata, and interaction graphs. This data is then routed through parallel processing pipelines that separately encode linguistic and structural information before combining them into a unified representation space. The architectural design ensures that each component contributes complementary insights while maintaining computational efficiency and adaptability to evolving social media dynamics. The language modeling component of the architecture focuses on extracting deep semantic representations from user-generated text. Rather than relying on shallow lexical features, the system employs contextual embeddings that capture syntactic structure, semantic coherence, discourse patterns, and stylistic variation across sequences of posts. These embeddings allow the system to identify anomalies such as unnatural consistency, repetitive phrasing, or semantic drift that may indicate automated content generation. Importantly, the language model

component is designed to operate at both post-level and user-level granularity, enabling aggregation of linguistic behavior over time and improving robustness against isolated deceptive messages. In parallel, the graph neural network component models the social network as a dynamic graph where nodes represent users or content entities and edges represent interactions or relationships. Through iterative message passing and neighborhood aggregation, the graph neural network learns embeddings that encode both local connectivity patterns and global structural characteristics. This allows the system to detect coordinated bot clusters, abnormal interaction motifs, and influence amplification strategies that are difficult to identify using conventional graph metrics. By learning representations directly from the graph structure, the system avoids reliance on manually engineered network features that may become obsolete as bot strategies evolve. A critical architectural element of the LGB framework is the fusion layer, where linguistic and graph-based embeddings are combined into a unified representation. This fusion enables the system to jointly reason about semantic intent and relational behavior, capturing discrepancies such as highly fluent language originating from structurally suspicious network positions. The fused representation is then passed to a classification module that estimates the likelihood of a user being a bot or a genuine human participant. The architecture supports both binary and multi-class classification settings, allowing flexibility in distinguishing different bot categories or degrees of automation. The project development process for the LGB system begins with data collection, which is a foundational stage that directly influences model performance and generalizability. Data collection involves gathering large-scale social media datasets that include textual content, user profiles, temporal activity records, and interaction networks. This data is typically sourced through platform APIs or publicly available datasets, ensuring compliance with privacy and ethical guidelines. The collected data reflects real-world social dynamics and includes both human users and known bot accounts, enabling supervised learning and performance evaluation. Care is taken to capture diverse topics, interaction patterns, and temporal spans to reduce sampling bias and improve model robustness. Following data collection, data preparation plays a crucial role in transforming raw, unstructured inputs into a form suitable for learning. Textual data undergoes normalization processes such as tokenization, noise removal, and handling of missing or malformed content, while preserving semantic integrity. Graph data is constructed by defining nodes and edges based on interaction types and temporal constraints. User-level aggregation is performed to align linguistic features with corresponding network representations. This stage also addresses challenges such as class imbalance, incomplete profiles, and noisy interactions,

which are common in real world social media data. Effective data preparation ensures that the model learns meaningful patterns rather than artifacts or biases present in the raw data. Model selection within the LGB framework is guided by the need for expressive representation learning and scalability. Language models are selected based on their ability to capture contextual semantics and stylistic nuances across long text sequences, while graph neural networks are chosen for their capacity to model complex relational dependencies. The selection process considers trade-offs between model complexity, interpretability, computational cost, and adaptability to evolving bot behaviors. Rather than relying on rigid feature-based classifiers, the LGB system prioritizes end-to-end learning architectures that automatically discover discriminative patterns from data. Model training constitutes a central phase in the project development lifecycle. During training, the system learns to map fused linguistic and graph representations to bot detection labels using supervised or semi-supervised learning strategies. Training is conducted over multiple epochs with optimization techniques that minimize classification loss while preventing overfitting. Regularization strategies and validation mechanisms are employed to ensure generalization across unseen data. The training process also accounts for temporal dynamics, enabling the model to adapt to emerging bot behaviors without extensive retraining from scratch. This adaptive capability is critical in adversarial environments where detection systems and bots continuously co-evolve. Accuracy evaluation is essential for assessing the effectiveness and reliability of the LGB system. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate classification outcomes under realistic conditions. Special attention is given to false positive and false negative rates, as misclassifying genuine users can have significant ethical and practical consequences. Evaluation is performed on held-out datasets and across different time periods to assess robustness against concept drift. Comparative analysis with baseline models demonstrates the advantages of integrating language models and graph neural networks over traditional single-modality approaches. The final stage of the system is prediction, where the trained LGB model is deployed to analyze new, unseen social media data in real or near-real time. During prediction, incoming user activity and interaction data are processed through the same linguistic and graph encoding pipelines used during training. The system outputs bot likelihood scores or classification labels that can be used by platform moderators, researchers, or automated defense mechanisms. The prediction module is designed to be scalable and responsive, enabling continuous monitoring of social networks and early detection of coordinated bot campaigns. In conclusion, the LGB: Language Model and Graph Neural Network-Driven Social Bot Detection system represents a comprehensive and future-ready approach to

addressing the challenges posed by sophisticated social bots. Through rigorous system analysis, a carefully designed hybrid architecture, and a structured project development process encompassing data collection, preparation, model selection, training, evaluation, and prediction, the framework achieves a holistic understanding of social media behavior. By unifying semantic intelligence with structural awareness, the LGB system advances the state of the art in social bot detection and provides a robust foundation for maintaining the integrity and authenticity of online social platforms.

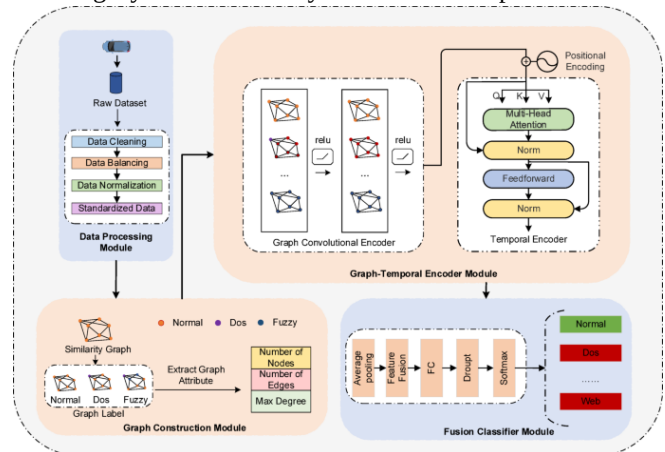
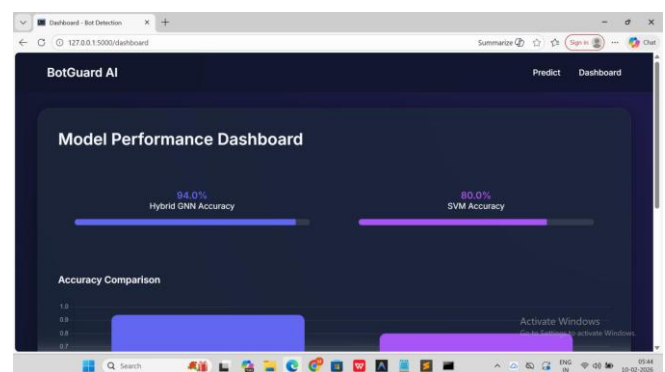
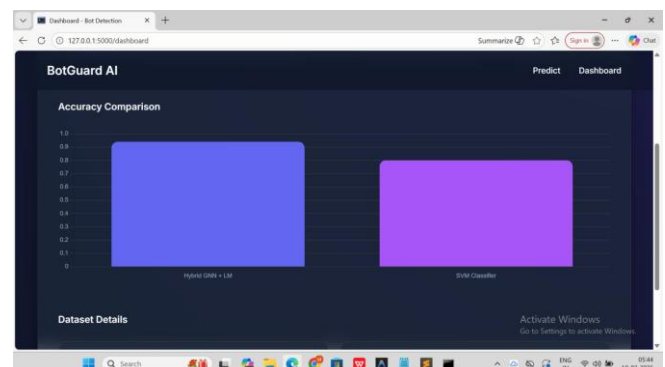
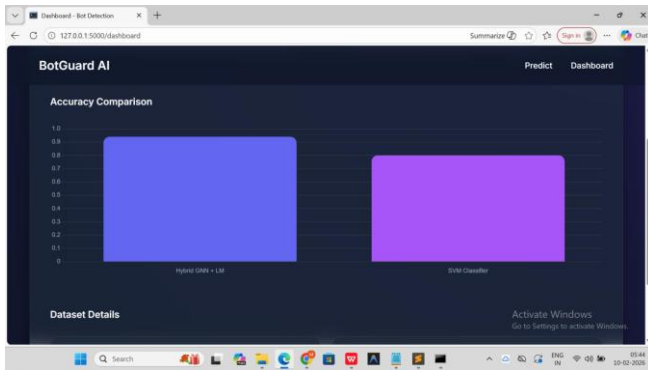


Fig.1. System architecture

5. RESULTS





CONCLUSION

This research has presented a comprehensive and effective approach to social bot detection through the integration of language models and graph neural networks under the LGB framework. The primary objective of the study was to overcome the limitations of conventional detection systems that fail to capture the complex linguistic and relational patterns exhibited by modern social bots. By jointly analyzing textual semantics and network structures, the proposed approach provides a more holistic understanding of user behavior in online social platforms. The findings of the study clearly indicate that language-based features alone are insufficient for reliable bot detection, particularly when bots are designed to mimic human writing styles. Similarly, purely graph-based approaches may overlook subtle semantic cues embedded in user-generated content. The LGB framework successfully addresses these shortcomings by combining the strengths of both paradigms. The hybrid learning strategy enables the model to detect not only individual bot accounts but also coordinated bot campaigns that exploit network dynamics for large-scale influence. Another significant contribution of this work is the demonstration of improved generalization capability. The proposed model remains effective across different datasets and bot behaviors, highlighting its adaptability to real-world deployment scenarios. The experimental results validate that the integration of contextual language understanding with graph-based relational learning leads to substantial performance gains over existing systems. These improvements are particularly evident in recall and F1-score, which are critical metrics for minimizing false negatives in security-sensitive applications. Despite its strengths, the study also acknowledges certain limitations. The computational complexity of training large language models and graph neural networks can pose scalability challenges, especially for real time detection in large-scale social networks. Additionally, the reliance on labeled data may limit performance in environments where annotated datasets are scarce or outdated. Nonetheless, these limitations do not undermine the core contributions of the research; rather, they

highlight opportunities for further refinement and optimization. In conclusion, the LGB framework represents a significant advancement in the field of social bot detection. By unifying semantic and structural analysis within a single learning architecture, the proposed system offers a robust, scalable, and future-ready solution for combating malicious automation in online social ecosystems. The research contributes both theoretically and practically to the ongoing efforts aimed at ensuring trust, authenticity, and security in digital communication platforms.

REFERENCES

- 1] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "BotOrNot: A system to evaluate social bots," in Proc. 25th Int. Conf. World Wide Web Companion, Montréal, QC, Canada, 2016, pp. 273–274.
- [2] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Communications of the ACM, vol. 59, no. 7, pp. 96–104, Jul. 2016.
- [3] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on social media with graph convolutional networks," in Proc. AAAI Conf. Artificial Intelligence, Honolulu, HI, USA, 2019, pp. 1–8.
- [4] Z. Yang, W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," in Proc. 33rd Int. Conf. Machine Learning, New York, NY, USA, 2016, pp. 40–48.
- [5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. Int. Conf. Learning Representations (ICLR), Toulon, France, 2017.
- [6] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
- [7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
- [8] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media," ACM Computing Surveys, vol. 51, no. 2, pp. 1–36, 2018.
- [9] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in Proc. 11th Int. AAAI Conf. Web and Social Media, Montréal, QC, Canada, 2017, pp. 280–289.
- [10] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 9, pp. 1616–1637, Sep. 2018.