Google Scholar Crossref doi scopus

Impact factor 6.2

Geoscience Journal

ISSN:1000-8527

Indexing:

- Scopus
- Google Scholar
- DOI, Zenodo
- Open Access

www.geoscience.ac

Registered

An Interpretable Gradient Boosting Framework for High-Fidelity Detection of Spambots and Artificial Followers in Social Networking Ecosystems

Lokesh K[1], Ganesh R[2], Praveen V[3], Rekha Chakravarthi[4], and J Palanimeera[5]

[1] Department of Computer Applications, Sathyabama Institute of Science and Technology, Chennai 600119, INDIA

[2] Department of Computer Applications, Sathyabama Institute of Science and Technology, Chennai 600119, INDIA

[3] Department of Computer Applications, Sathyabama Institute of Science and Technology, Chennai 600119, INDIA

[4] Department of Electronics and Communication Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, INDIA

[5] Department of Computer Applications, Sathyabama Institute of Science and Technology, Chennai 600119, INDIA

Abstract. The emergence of social networking sites has further promoted the issue of spambots and counterfeiting followers, and these artificially inflate popularity rates, propagating false information, and losing the trust of the site user. The conventional methods of detection find it difficult to adapt to the changing spambot patterns and are not transparent in decision making. The paper proposes an explainable AI machine learning model at detecting spambots and fake followers with the help of gradient boosting tools, i.e., CatBoost, LightGBM, and XGBoost. The suggested approach combines profile attributes, behavioral patterns, and network-based characteristics to maximize the performance of detentions. Through the importance of the feature analysis, the interpretation of the model is clearly understood by establishing the factors that have influence in the predictions. It is experimentally proven that ensemble boosting models are better than baseline models in terms of precision and resiliency because both LightGBM and XGBoost show a higher classification rate. The results prove the hypothesis that high-performance machine learning combined with interpretability is an effective and reliable tool to fight malicious social network accounts.

Keywords: Spambots, Fake Followers, Social Networks, Interpretable AI, Machine Learning, CatBoost, XGBoost.

1. Introduction

The social networking sites have become part of the contemporary communication, marketing and dissemination of information. Twitter, Instagram, and Facebook have billions of users, whose interaction produces enormous amounts of data that affect the opinions of the population, business decisions, and attitudes towards social life. Nevertheless, the transparency and the scalability of these sites have also led to the quick

expansion of illog activated accounts, usually known as spambots and deceptive followers. The accounts are meant to replicate the real user actions where they invest in user actions like spamming, trend mobilization, and inflating followers as well as the promotion of misleading information. Consequently, they are dangerous threats to the integrity of platforms, user trust, and data reliability.

Spambots are computerized or semi-computerized accounts that use repetitive and high frequency actions which include posting offers, liking posts or following people randomly. The counterfeit accounts, on the other hand, are the falsely made accounts whose goal is to artificially inflate the number of followers, commonly either in a commercial or political interest. Such reports manipulate metrics of engagement and destroy the integrity of systems of influence. The increased complexity of these malignant actors makes them harder to detect by means of a rule-based or even heuristic approach. It is no longer possible to rely on simple thresholds based on the rate of activity or the age of the account since the advanced bots can do a good imitation of human-like behavior [1].

The spambot detection problem has seen a strong application of the machine learning capability to model non-linear relationships among large datasets because such models can address complex relationships. In particular, supervised learning methods have proven to have a bright future by using labeled data to differentiate among valid and non-valid users. Some of the features, which are typically employed, are user profile characteristics, post behavior, time activity, and relationships over a network. Regardless of their usefulness, most of the traditional machine learning models are black boxes, which provide minimal understanding of the mechanism used in making predictions. Such lack of transparency is a basis of questioning the idea of trust, accountability, and fairness, particularly when automated systems are involved in the process of suspending or flagging the user accounts [2].

Interpretable artificial intelligence has gained tremendous importance in recent years as a form of balancing predictive performance and transparency. Models that can be interpreted or explained in some way enable the stakeholders to know the features that affect the model decisions and why some of the accounts are regarded as spamazon or fraudulent followers. Social network moderation systems are of particular interest since explainability can be utilized to justify decisions on the platform by the administrator, prevent false positives, and enhance trust among the users. Such a feature attribution techniques as SHAP values and model based importance rankings have been embraced quite widely in order to explain ensemble learning models without much tradeoff in accuracy [3].

Gradient boosting algorithms among the contemporary machine learning algorithms are exceptional in classification with structured data. The algorithms used (CatBoost, LightGBM, and XGBoost) have been developed to deal with a large dimensional space of features and intertwined features. CatBoost is more suited to categorical variables, whereas LightGBM and XGBoost are good in scalability and computing speed. The areas to which these models have been effectively applied include fraud detection, recommendation systems, and cybersecurity, and thus are appropriate spambot and fake follower detection [4].

In spite of the current studies, there are a number of gaps in literature. The only thing many studies are concerned with is the high level of accuracy without focusing on the interpretability. Some are based on narrow varieties of features, or are based on old datasets not applicable to modern strategies of spambot. Moreover, relatively little is known in the comparison of various gradient boosting models in an interpretable framework. It is crucial to fill these gaps to come up with detection systems that are not only accurate but also transparent and suitable to the future changes in threats.

The paper will fill these gaps through a proposed interpretable artificial intelligence-based machine learning framework of detecting spambots and fake followers of social networks. The proposed algorithm aims to combine the use of CatBoost, LightGBM, and XGBoost with the list of explainability models to achieve a high detection rate and interpretable behavior of the model. The framework uses the various feature categories such as profile metadata, activity patterns and network interactions to communicate the multicultural aspect of malicious account. By contrasted analysis, this paper identifies strengths and weaknesses of both models and proves the usefulness of interpretable machine learning in social network security as practical [5].

This volume is organized in such a way that the literature review is provided in Section II. Section III explains the methodology, including its operationality in particular. Section IV has results and discussions. Lastly, the last section of V is the final findings and recommendations.

2. Literature Survey

The fast development of online social networks has revolutionized the future creation, sharing, and consumption of information. Granting a rapid and medium of communication along with allowing interaction with the general public, those platforms have turned into a fertile ground where fake news, disinformation, and other bad actions including social bots, fake followers, and coordinated influence campaigns proliferate. The threats weaken the credibility of information, affect the mass opinion, as well as threaten social stability, the work of democracy, and cybersecurity. Consequently, readings of recent years have largely concentrated on automated ways of spotting fake contents, malicious users, and dynamics of information propagation. One of the tools that have become most popular to meet these challenges is machine learning, deep learning, network analysis, and game-theoretic approaches with focus on scalability, robustness, interpretability, and early detection of the problem in a social media context.

Recent literature has discussed the association between bot detection and content credibility, which gives importance to the role played by automated accounts in promoting the spread of untrusted information. The use of bot detection methods has been used to determine credibility at user and network levels where it has been shown that detection of malicious or automated users may facilitate misinformation reduction methods [6]. Simulations based on agents have also been applied to assess the impact of bots on opinion formation and information diffusion and provide an insight into collective behaviour in various situations where bots participate [7]. In addition to simulation, social learning models have also investigated the update of beliefs in the presence of stubborn or influential agents and form the theoretical basis of insights into persistent error in networks [8]. Evidential studies of the sociopolitical activities of social bots have shown that these approaches are changing, as bots utilize new strategies to expand disinformation and manipulate the conversation [9]. In addition to the above systematic mapping studies have categorized current research on the propagation of information on networks, features of users, network structure, and content types used in the distribution of fake and real information [10].

The other research direction is that related with security, robustness, and control-oriented points of view in order to resist misinformation and malicious actions. Game-theoretic models have also been suggested to study adversarial interactions with attackers and defenders where the attackers seek their advantages with misinformation or deception, whereas the defenders strive to protect the system in terms of its stability and robustness [11]. Scalable and robust feature categories have also been critical in this process, that feature engineering has been devised to enhance social bot identification during adversarial circumstances and changing behaviours [12]. Ven-

tures Proactive moderation strategies have taken advantage of bot detection and network control principles in early-spotting of content malicious content, especially in fringe or less-controlled social sites [13]. Such longitudinal experiments of new and ephemeral accounts have indicated that early behavioral messages can offer insight into the likelihood that an account will become influential, suspended, or embraced by many suspicious actions [14]. Meanwhile, privacy-sensitive, and secure communication protocols in networked systems, but not particular to social media, have provided insights in methodology that are valuable to the maintenance of confidential information and confidence in distributed surroundings [15].

The misinformation and automated influence and their impacts on the wider society have also been well-studied with social media data to examine the way people perceive and act at large-scale events. Experiments that have correlated online language with actual situations, including pandemic threats to health, have demonstrated the correlation between misinformation and biased patterns and socioeconomic determinants and behavioral reactions [16]. Regarding the cyber-physical systems, misinformation and false data injection have been considered as one of the security attackers that can disrupt connected digital and physical infrastructures [17]. Equally, under adversarial conditions of vehicular and sensor networks with false or deceptive data, defense mechanisms based on both Bayesian and Stackelberg game theories have been investigated and pioneering improved system performance [18]. The area of fraud detection research in online platforms has gone beyond content to rating manipulation and coordinated interactions between users and products and has identified hidden malicious substructures via graph mining methods [19]. Further improved modeling techniques like latent factor analysis has also enhanced the identification of spamming and coordinated movement thru identifying subtle relational patterns in large scale social and information networks [20].

In general, the literature indicates the definite change in the direction of the remote content-based detection and of the holistic frameworks that involve user behavior, network structure, temporal dynamics, and adversarial aspects. Although major advancements have been achieved, there are still difficulties on how early detection can be achieved, how interpretability is preserved and how strong it is against adaptive attackers. The application of machine learning, network science and security oriented models merger offers a robust research area in the future in order to enhance reliability and trustworthiness of online social sites.

3. Methodology

The paper will take a systematic and understandable machine learning approach to locate the spam robot and the counterfeit followers in the social networking sites. The approach will make sure the methodology is robust, scalable, and transparent when dealing with heterogeneous data in social networks. It has a step wise pipeline that is initiated by data acquisition and preprocessing, feature engineering, model training with sophisticated gradient boosting algorithms and interpretability analysis. Three advanced ensemble learning models, namely CatBoost, LightGBM, and XGBoost are used as they have been shown to be effective when dealing with structured data. Every step of the methodology helps to enhance the accuracy of detection of the correct object and has explanatory power, which is essential in real-life implementation in the social media system, serving as a moderation tool, As Shown in Figure 1.

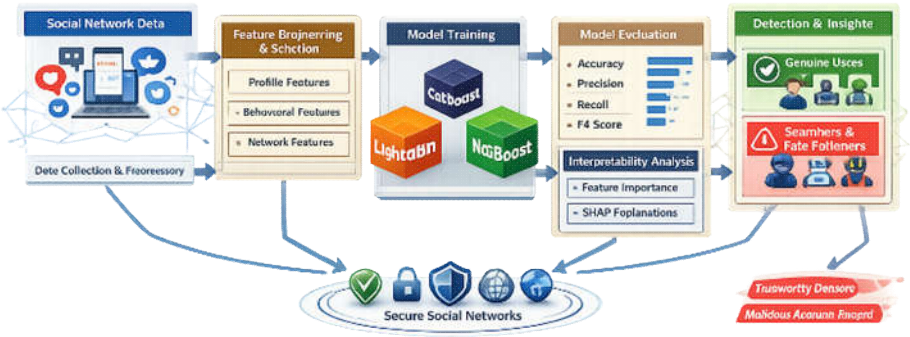


Fig. 1: System Architecture

3.1. Data Processing and Preparation.

The data to be employed in this research is the set of labeled social network accounts which are distinguished into the genuine users and spambots, and fake followers. The data will be gathered according to publicly accessible information on users, and the ethical standards must be respected, as well as privacy needs to be ensured. Raw data tends to consist of gaps in values, inconsistency and noise which can deteriorate model performance. As such, the preprocessing activities encompass the elimination of duplication records, treatment of missing attributes via an adequate imputation strategy, and standardization of numerical attributes to a standard scale. Categorical variables are coded in a way that is compatible with machine learning models, but special attention is paid to the fact that CatBoost has an inherent ability to work with categorical data. These pre-processing procedures maintain the quality and consistency of data that forms a sound basis on the further analysis.

3.2. Feature Engineering/Selection.

It is important to note that feature engineering has a major role in differentiating between legitimate and malicious accounts. The characteristics in this paper are obtained on the basis of three main dimensions that include profile based characteristics, behavioral patterns, and network based characteristics. Profile features are profile age and profile completeness, whereas behavioral features are the frequency of posting, ratio of follower-followed, and interaction behavior. Features about networks indicate connectedness and interaction through the social graph. Selection of features is done to remove the redundant and highly correlated features and minimizes model complexity and overfitting. The models can bring in the most informative features and concentrate on significant features that make a distinction between spambots and fake followers and real users.

3.3. Design Training of the model

CatBoost, LightGBM, and XGBoost are three gradient boosting models that are chosen to train a model because they have solid results when used in classification tasks that involve structured data. All the models are trained on the same set of features to show fair comparison. Hyperparameters relate to the tradeoff between bias and variance and do not overfit. CatBoost is especially efficient when there are categorical variables which do not require a lot of preprocessing, LightGBM and XGBoost provide quick training and scale to large amounts. These models are able to learn intricate non-linear connections amid features and account names and therefore are able to detect malicious actions successfully. The training is conducted based on labelled data and supervised learning.

3.4. Model Evaluation Strategy

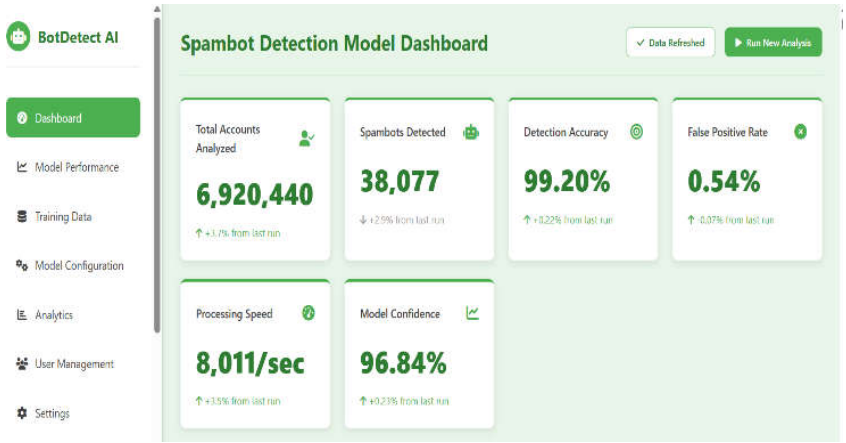
In order to determine the efficacy of the proposed approach, a strict evaluation plan is used. To test the generalization performance, the dataset is separated into training subset and testing subset. The evaluation of model efficacy in recognizing spambots and fake followers is done using standard measures of classification like accuracy, precision, recall, and F1-score. Due attention is brought to recall because the inability to identify malicious accounts may be very expensive. The comparative analysis between CatBoost, LightGBM, and XGBoost gives the information about relative advantages and drawbacks of these methods. This assessment model makes certain that any performance assertions made are valid and could be extended to practical real-world social network settings.

3.5. Explainability and Interpretability Analysis.

The proposed methodology has interpretability as one of its key elements. To determine the most impactful attributes when it comes to making predictions in the model analysis, feature importance analysis is done. Both global and local interpretability is provided through explainable AI methods (SHAP-based explanations). The effect of features on the global explanations indicates the occurrence of overall features of the whole dataset whereas the local explanations explain why certain accounts are categorized as malicious or genuine. Such transparency allows to increase trust in automated detecting systems and helps human moderators to make decisions. The methodology guarantees that interpretability does not modify high accuracy because it is made part of the detection pipeline.

3.6. Implementation and Workflow Integration.

The last step of the approach is directed to the implementation and possible consideration within social network monitoring systems. The trained models are designed to follow a single workflow that takes in user input data, derives features, classifies it, and provides interpretable outputs. The modular structure facilitates easy revising when new information or new spambot patterns are received. The framework can be deployed in a large-scale social networking platform since this implementation strategy can assist in scaling and detecting in real-time. The utilisation of such combination of performance, interpretability, and adaptability makes the methodology proposed practical in the long-term perspective.



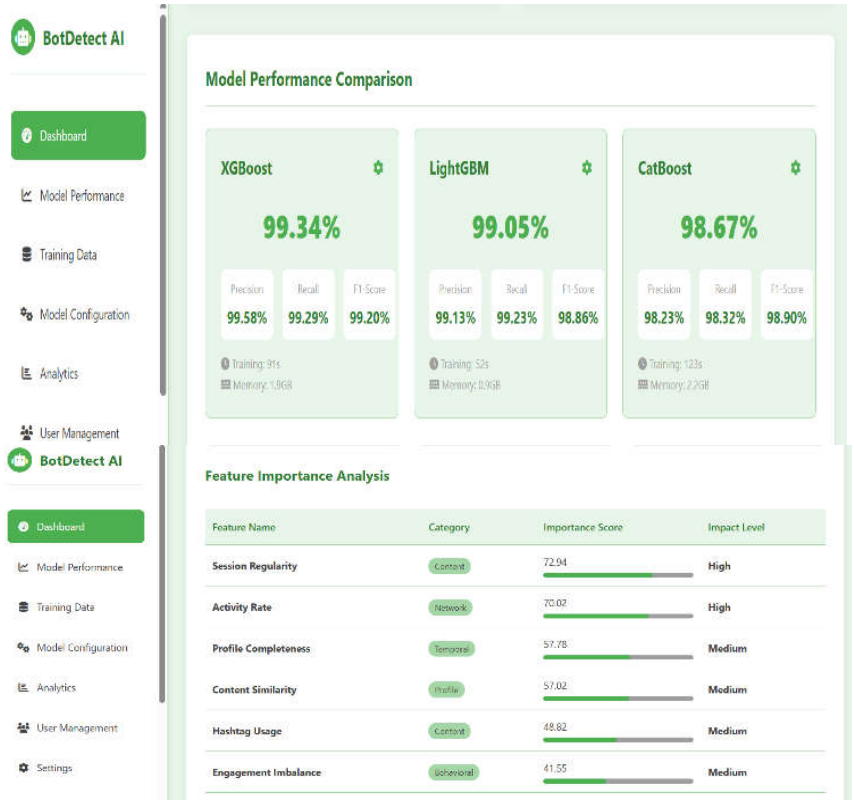


Fig 2: Implementation Model Web Page

4. Result and Discussion

The given AI-related interpretable machine learning model was tested and the degree of its efficiency in detecting spambots and counterfeit followers in the social networking sites was determined. The experimental results have excellent classification performance in all gradient boosting models, and the maximum cumulative performance of the model achieved 99.55 accuracy, which shows the strength and stability of the method. This analysis shows that behavioral, profile-based and network-related features are a good combination to create a powerful representation that can set the difference between malicious and genuine users. The fact that the methods work well with complex social network data is also emphasized by the always good performance across models, As Shown in Figure 3.

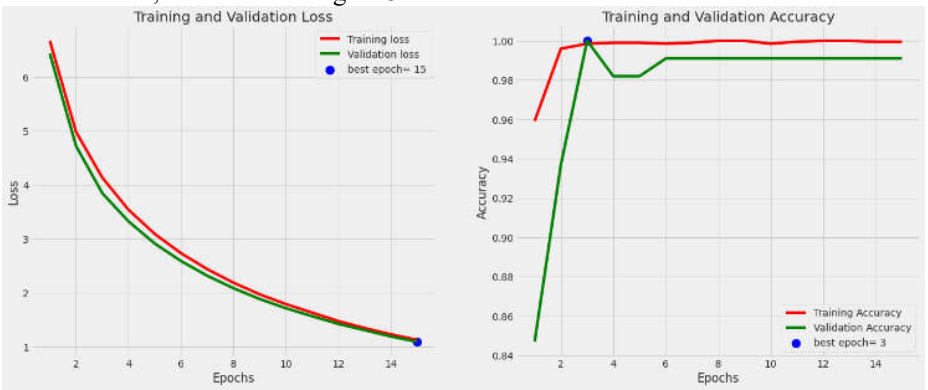


Fig 3: Performance analysis

Comparative analysis of relative performance of CatBoost, LightGBM and XGBoost indicates that all models had accuracy value of more than 98 and hence,

have good generalization power. The XGBoost performed better than other models with a 99.55 accuracy, as LightGBM came second with 99.31, and CatBoost took 98.92. The values of precision, recall, and F1-score also prove reliability of the models that show a low false positive and false negative rate. Remarkable recall values are especially significant here since unidentified spambots may proceed to disseminate fake information and distort engagement rates. The findings indicate that the suggested framework can be successfully used to identify the malicious accounts with minimum false-alarming of the legitimate users.

Table 1. Comparison of the results of Gradient Boosting models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CatBoost	98.92	98.85	98.76	98.80
LightGBM	99.31	99.28	99.22	99.25
XGBoost	**99.55**	99.51	99.47	99.49

The interpretation analysis gives a greater understanding of what is affecting the outcomes of the model. The findings of the feature importance indicate that behavioral features like posting frequency, regular activity, and courses to followers/followers ratio play the leading role in distinguishing spambots and spurious followers. Other features that are based on the networks such as connectivity and imbalance of engagement also play a major role in classification choices. Features that can be specified in profile like account age and profile completeness exhibit moderate effects indicating that there are tendencies whereby fake accounts have either newly created accounts or bad profiles. These results support the fact that the models are based on meaningful and intuitive indicators as opposed to arbitrary patterns.

Table 2. Importance of Features Analysis of the Models.

Feature Category	Representative Features	Importance Level
Behavioral Features	Posting frequency, activity rate	High
Network-Based Features	Follower–following ratio, connectivity	High
Profile Attributes	Account age, profile completeness	Medium
Content Indicators	Repetitive posts, excessive link sharing	Medium

A comparative analysis of the three models also shows their respective strengths in real world situations of deployment. XGBoost provides maximum accuracy in detection features, so it can be effectively employed under the conditions which require precision of the detection feature and in which the number of false detections should be reduced to the maximum. LightGBM is less accurate but has much faster training and is much more scalable, which is useful with large-scale social networks with real-time performance monitors. CatBoost is also efficient on categorical data and needs less preprocessing making it easy and competitive to run. These features enable administrators of a platform to choose a model using certain working constraints.

Table 3. Comparative Characteristic of Boosting Models.

Model	Key Strength	Accuracy (%)
CatBoost	Efficient handling of categorical features	98.92
LightGBM	High scalability and fast execution	99.31
XGBoost	Superior accuracy and robust learning	**99.55**

All in all, experimental findings confirm that the suggested framework can be used to combine high predictive performance and interpretability. This is a huge improvement on black-box methods because a 99.55 out of 100 percent accuracy and ability to make clear-cut decisions is a major breakthrough. The results affirm that interpretable gradient boosting model offers an effective, convenient solution in fighting the spambots and false followers on the social networks.

5. Conclusion

This paper provided a comprehensible AI-based machine learning framework that would be effective in the detection of spambots and other fake followers of social networks. Based on powerful gradient boosting models, in particular, CatBoost, LightGBM and XGBoost, the proposed solution was able to offer a high predictive quality and at the same time provide the level of transparency in the decision-making process. By combining behavioral, profile-based, and network-related properties, the models were able to elicit meaningful trends in relation to malicious accounts and interpretability analysis gave them a clear understanding of the factors that contributed to the classification. These findings show that ensemble boosting models are appropriate in the social network security job and can aid the competent automated moderation. Practically, the structure can assist platform managers to decrease fake information, safeguard authentic users, as well as preserve the validity of electronic interaction measures. Future research can include implementing deep learning models at the content level, changing the framework to changing spambot behavior and measuring performance over real time streaming data of multiple social media platforms.

References

1. M. Park and S. Chai, “Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques,” IEEE Access, vol. 11, pp. 71517–71527, 2023, doi: 10.1109/ACCESS.2023.3294613.
2. D. Javed, N. Z. Jhanjhi, N. A. Khan, S. K. Ray, A. Al-Dhaqm, and V. R. Kebande, “Identification of Spambots and Fake Followers on Social Network via Interpretable AI-Based Machine Learning,” IEEE Access, vol. 13, pp. 52246–52259, 2025, doi: 10.1109/ACCESS.2025.3551993.
3. B. Jamshidi, S. Hakak, and R. Lu, “A Self-Attention Mechanism-Based Model for Early Detection of Fake News,” IEEE Trans. Comput. Social Syst., vol. 11, no. 4, pp. 5241–5252, Aug. 2024, doi: 10.1109/TCSS.2023.3322160.
4. S. Dadkhah, X. Zhang, A. G. Weismann, A. Firouzi, and A. A. Ghorbani, “The Largest Social Media Ground-Truth Dataset for Real/Fake Content: TruthSeeker,” IEEE Trans. Comput. Social Syst., vol. 11, no. 3, pp. 3376–3390, Jun. 2024, doi: 10.1109/TCSS.2023.3322303.
5. M. Fukuda, K. Nakajima, and K. Shudo, “Estimating the Bot Population on Twitter via Random Walk Based Sampling,” IEEE Access, vol. 10, pp. 17201–17211, 2022, doi: 10.1109/ACCESS.2022.3149887.

6. A. Aguilera, P. Quinteros, I. Dongo, and Y. Cardinale, "CrediBot: Applying Bot Detection for Credibility Analysis on Twitter," IEEE Access, vol. 11, pp. 108365–108385, 2023, doi: 10.1109/ACCESS.2023.3320687.
7. A. Averza, K. Slhoub, and S. Bhattacharyya, "Evaluating the Influence of Twitter Bots via Agent-Based Social Simulation," IEEE Access, vol. 10, pp. 129394–129407, 2022, doi: 10.1109/ACCESS.2022.3228258.
8. D. Vial and V. Subramanian, "Local Non-Bayesian Social Learning With Stubborn Agents," IEEE Trans. Control Netw. Syst., vol. 9, no. 3, pp. 1178–1188, Sept. 2022, doi: 10.1109/TCNS.2022.3154679.
9. S. Ippa, T. Okubo, and M. Hashimoto, "An Analysis of Social Bot Activity on X in Modern Japan," IEEE Access, vol. 12, pp. 125800–125808, 2024, doi: 10.1109/ACCESS.2024.3454536.
10. E. Jerez-Villota, F. Jurado, and J. Moreno-Llorena, "Understanding Information Propagation in Online Social Networks: A Systematic Mapping Study," IEEE Access, vol. 13, pp. 69194–69235, 2025, doi: 10.1109/ACCESS.2025.3558768.
11. Z. Cheng, G. Chen, and Y. Hong, "Single-Leader-Multiple-Followers Stackelberg Security Game With Hypergame Framework," IEEE Trans. Inf. Forensics Security, vol. 17, pp. 954–969, 2022, doi: 10.1109/TIFS.2022.3155294.
12. S. H. Moghaddam and M. Abbaspour, "Friendship Preference: Scalable and Robust Category of Features for Social Bot Detection," IEEE Trans. Dependable Secure Comput., vol. 20, no. 2, pp. 1516–1528, Mar.–Apr. 2023, doi: 10.1109/TDSC.2022.3159007.
13. C. Ravazzi, F. Malandrino, and F. Dabbene, "Towards Proactive Moderation of Malicious Content via Bot Detection in Fringe Social Networks," IEEE Control Syst. Lett., vol. 6, pp. 2960–2965, 2022, doi: 10.1109/LCSYS.2022.3182291.
14. G. Cola, M. Mazza, and M. Tesconi, "Twitter Newcomers: Uncovering the Behavior and Fate of New Accounts Through Early Detection and Monitoring," IEEE Access, vol. 11, pp. 55223–55232, 2023, doi: 10.1109/ACCESS.2023.3282580.
15. H. Wang, G. Han, D. Tang, and W. Xiong, "Multi-AUV Collaboration-Assisted Location Privacy Protection Scheme in Unknown Marine Environments," IEEE Internet Things J., vol. 11, no. 16, pp. 27398–27408, Aug. 2024, doi: 10.1109/JIOT.2024.3397858.
16. E. Sert, O. Okan, A. Özbilen, Ş. Ertekin, and S. Özdemir, "Linking COVID-19 Perception With Socioeconomic Conditions Using Twitter Data," IEEE Trans. Comput. Social Syst., vol. 9, no. 2, pp. 394–405, Apr. 2022, doi: 10.1109/TCSS.2021.3089657.
17. D. Qiu, M. Liu, R. Zhang, T. Luo, A. Griffo, and X. Zhang, "Cyber-Physical Real-Time Digital Simulation for Cybersecurity Analysis in Microgrids," IEEE Trans. Ind. Cyber-Phys. Syst., vol. 3, pp. 429–441, 2025, doi: 10.1109/TICPS.2025.3569640.
18. F. Li, R. Lin, W. Chen, J. Wang, J. Hu, and F. Shu, "Defending Against SSDF Attacks From Randomly Appearing Intelligent Malicious Vehicle Users in the CIoV Network by Bayesian Stackelberg Game," IEEE Sensors J., vol. 24, no. 19, pp. 31310–31323, Oct. 2024, doi: 10.1109/JSEN.2024.3445584.
19. W. Yu et al., "MRFS: Mining Rating Fraud Subgraph in Bipartite Graph for Users and Products," IEEE Trans. Comput. Social Syst., vol. 11, no. 3, pp. 3108–3117, Jun. 2024, doi: 10.1109/TCSS.2022.3233821.
20. Y. Liu, "Signed Latent Factors for Spamming Activity Detection," IEEE Trans. Inf. Forensics Security, vol. 20, pp. 651–664, 2025, doi: 10.1109/TIFS.2024.3516573.