International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research Paper

Smart Detection of Automated Social Media Accounts Using Feature Optimization and Ensemble Learning Techniques

K Haritha[1], D V Sushma Charitha[2], R R Shantha Spandana[3]

[1]P.G Scholar, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: karthikharitha34@gmail.com, ORC-ID: https://orcid.org/0009-0006-9669-9590

[2] M.Tech Scholar, Department of CSE, Sri Venkateswara University College of Engineering ,Tirupati,
E-mail: dv.charitha@gmail.com , ORC-ID: https://orcid.org/0009-0001-0759-3254

[3]Assistant Professor, Department of MCA, Sri Venkatesa Perumal College of Engineering & Technology, Puttur,
E-mail: shanthaspandana@gmail.com, ORC-ID: https://orcid.org/0009-0003-4236-1250

Abstract: Social networking services like Twitter facilitate extensive human contact but are increasingly infiltrated by automated accounts that emulate human behavior, disseminating misinformation and influencing public opinion. Identifying such spambots is essential for preserving information integrity; however, numerous traditional methods depend on opaque, black-box models, which restrict transparency and interpretability. This research utilizes the Cresci-15 and Cresci-17 datasets to examine interpretable machine learning methods for detecting spambots and fraudulent followers. Both feature-based and text-based data are employed, using preprocessing techniques such as normalization, tokenization, and the elimination of extraneous content. Recursive Feature Elimination (RFE) decreases feature dimensionality, whereas resampling techniques like SMOTE and SMOTEENN rectify class imbalance. A variety of machine learning methods, such as Decision Tree, Random Forest, SVM, XGBoost, AdaBoost, Stacking Classifier, and Voting Classifier, are assessed. The results indicate that the Stacking Classifier attains exceptional performance, achieving 99.9% accuracy on the Cresci-15 dataset and 99.5% accuracy on the Cresci-17 dataset. Moreover, explainable AI techniques like LIME and SHAP offer explicit insights into feature significance, hence improving model transparency and facilitating informed decision-making. These findings underscore the efficacy of integrating feature selection, sophisticated resampling, and ensemble learning methodologies with interpretable techniques for the reliable identification of automated accounts in social networks.

"Index Terms: *Interpretable AI, social network, bot detection, fake followers, spambots."*

1. INTRODUCTION

Social networks have become the primary conduit for information diffusion in the contemporary digital age. X (previously Twitter) has emerged as one of the most prominent and impactful platforms, facilitating real-time connections and interactions among millions of active users [1]. Its extensive social and economic influence has, however, drawn evil entities that exploit its open framework to distort public perception and disseminate disinformation. Automated programs, referred to as bots, are prevalent technologies utilized for this goal. Although several bots provide advantages by producing valuable content like instructional blogs and news updates, malignant bots are engineered to spread spam, misinformation, and detrimental material [2].

Identifying these rogue identities has emerged as a significant difficulty in maintaining the integrity of online conversation. Bot detection methodologies frequently depend on user metadata, behavioral characteristics, timestamps, and network architectures [3]. However, the process of feature engineering—designing and extracting these attributes—requires significant work and specialized knowledge. Furthermore, bots are continually advancing, employing adaptive tactics that replicate human behavior, so rendering them progressively more challenging to identify. Malicious bots expedite the widespread dissemination of misinformation, such as fake news and hate speech, by strategically engaging with prominent accounts [4].

The ecosystem of automated manipulation is further compounded by the presence of botnets and Sybil accounts. A botnet denotes a network of synchronized bots doing particular nefarious activities, whereas Sybil accounts signify counterfeit identities employed to manipulate genuine dialogues. Such techniques undermine authentic engagement and elevate low-credibility information, prompting significant worries about trust and reliability across social media ecosystems.

In light of these risks, machine learning (ML) has been extensively utilized in fields such as sports analytics [7], sentiment analysis [8], [9], and the detection of fake news [10]. In recent years, machine learning-based methodologies have been used to

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

social bot detection, providing automated solutions that surpass conventional heuristic and network-based techniques. Nonetheless, several machine learning-based detection systems continue to function as "black-box" models, offering restricted interpretability. The absence of transparency complicates the ability of researchers and practitioners to discern whether predictions are founded on significant patterns or merely overfitted noise.

To mitigate these constraints, interpretable machine learning (XAI) is progressively incorporated into bot detection systems. XAI augments transparency by elucidating the impact of individual attributes on model predictions, hence enhancing accountability and confidence. XAI-based frameworks enhance the accuracy and interpretability of algorithmic judgments, facilitating more reliable and trustworthy detection of bots on X and comparable social networks.

2. LITERATURE REVIEW

The identification of bots on social networks has garnered considerable focus in recent years because of the dynamic and evolving characteristics of fraudulent accounts that interfere with online communication and distort information ecosystems. Researchers have progressively investigated sophisticated machine learning techniques, graph-based methodologies, and interpretable frameworks to improve detection precision and generalizability. Mbona and Eloff [11] underscored the necessity of categorizing bots as both malevolent and benign, emphasizing that not all bots fulfill destructive roles. They presented a semi-supervised machine learning methodology that utilizes a limited amount of labeled data in conjunction with a substantial quantity of unlabeled data, effectively mitigating the deficiency of annotated datasets. Their research indicated that semi-supervised learning serves as an effective solution in contexts with elevated labeling costs, enhancing the distinction between malicious bots and benign automated accounts, such as those disseminating news updates.

Yang, Varol, Hui, and Menczer [12] addressed the challenges of scalability and generalizability in bot identification. They proposed a data selection technique to enhance the training process of machine learning models. By meticulously selecting representative training data, their methodology guarantees that models excel across diverse datasets and are not confined to certain contexts or restricted scenarios. This contribution is substantial due to the primary problem in bot detection, which is the flexibility of models to diverse datasets and the evolution of bot behaviors. Their research demonstrated that enhancing training data selection is equally vital as optimizing algorithmic models.

Lopez-Joya, Diaz-Garcia, Ruiz, and Martin-Bautista [13] conducted a comprehensive examination of bots utilizing generative AI technology. They analyzed the structure of these bots and found emerging trends and issues. In contrast to conventional spambots, generative AI-driven bots may generate very realistic and contextually pertinent language, rendering them more difficult to differentiate from human users. Their research emphasized the pressing necessity to modify detection systems to address emerging categories of bots that employ massive language models and sophisticated natural language creation methods. This underscores how advancing AI technology enable both defenders and enemies.

Pham [14] suggested an integrated methodology that combines a pre-trained auto-encoder with a graph neural network (GNN) for the detection of social bots. The model utilizes the representational capabilities of auto-encoders to capture latent behavioral characteristics and combines them with GNNs to describe social interactions. This dual methodology enhances generalizability and enables the detection system to adjust to various bot types across networks. Their research exhibited enhanced efficacy relative to traditional supervised learning methods, indicating that hybrid deep learning models can encapsulate both structural and behavioral intricacies of social media profiles.

To enhance user trust in online platforms, Terumalasetti and Reeja [15] suggested an extensive detection framework utilizing multi-dimensional analytics. Their research concentrated on identifying harmful profiles by integrating multiple aspects, including content characteristics, temporal patterns, and relational factors. The framework, by including several analytical views, strengthened resilience against adaptive spambots and fostered increased user confidence. This method underscores the necessity for comprehensive models that incorporate various feature categories instead of depending on a singular analytical dimension.

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

Previous research conducted by Paudel, Nguyen, Hatua, and Sung [16] examined the emerging phenomenon of social bots on Twitter through the application of complex network analysis methodologies. Their research demonstrated that bot methods have evolved to become more complex, with bots emulating human interaction patterns and integrating themselves extensively throughout networks. Through empirical investigation, they revealed that the structural aspects of networks, including clustering coefficients and centrality metrics, could serve as valuable indications for bot detection. Their input was essential in comprehending the evolution of bot activities and the constraints of conventional detection technologies.

Aljabri, Zagrouba, Shaahid, Alnasser, Saleh, and Alomari [17] performed a literature study on machine learning-based social bot detection, grounded in extensive analyses. They assessed current methodology, emphasizing strengths and weaknesses in supervised, unsupervised, and deep learning techniques. The analysis highlighted significant concerns, including excessive dependence on feature engineering, scalability issues, and the absence of interpretability in black-box models. Their findings suggested that future advancements in bot detection rely on the creation of models that harmonize performance with transparency and adaptability.

Wu, Ye, and Mou [18] presented Botshape, an innovative detection methodology that utilizes behavioral pattern analysis. Their system transcends static account metadata by modeling dynamic behaviors, including interaction frequency and temporal posting patterns. By identifying these behavioral subtleties, Botshape attained enhanced precision in differentiating between bots and humans, especially in situations where bots endeavored to replicate human timing and interaction patterns. This method illustrated the increasing significance of behavioral modeling in bot identification.

Bebensee, Nazarov, and Zhang [19] investigated the application of node neighborhoods and egograph topology to improve bot identification in social graphs. Their approach utilized structural information from local neighborhoods in graph data, providing an enhanced means to discern relationship distinctions between bots and people. Their model, centered on egograph topology, yielded insights into micro-level social structures and enhanced detection performance without the need for substantial handcrafted features. This method demonstrates how graph-based learning can offer more substantial contextual indicators for detecting unusual actions.

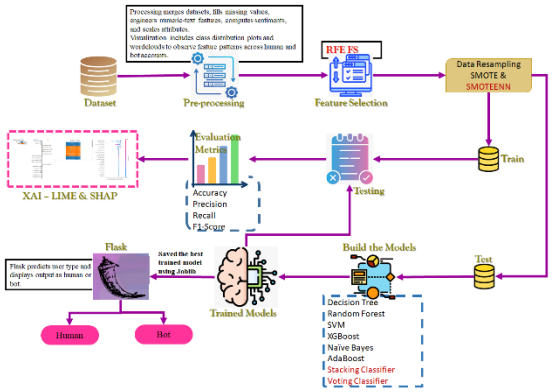
Finally, Dimitriadis, Dialektakis, and Vakali [20] introduced Caleb, a conditional adversarial learning framework aimed at enhancing bot identification. Their strategy employed adversarial learning to enhance resilience against adaptive bot strategies designed to circumvent detection. The system enabled the detection model to learn from hostile instances, hence enhancing generalization and fortitude against emerging threats. Their contribution is significant in tackling the adversarial dynamics of bot detection, where malevolent entities persistently enhance tactics to circumvent automated defenses.

3. MATERIALS AND METHODS

The proposed method presents a robust and interpretable machine learning approach for identifying spambots and counterfeit followers on social networks by amalgamating feature-based and text-based data from the Cresci-15 and Cresci-17 datasets. Data preprocessing encompasses the management of missing data, encoding of categorical variables, standardization of numerical characteristics, text cleaning (including the removal of HTML tags, URLs, and non-printable characters), normalization, tokenization, and TF-IDF representation. SMOTE and SMOTEENN are utilized to rectify class imbalance. Predictive models are constructed utilizing Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, Naïve Bayes, and AdaBoost, augmented by Stacking and Voting Classifiers. Recursive Feature Elimination decreases dimensionality, yet LIME and SHAP guarantee interpretability. A Flask-based interface facilitates real-time detection and interactive study of model decisions.

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research

Paper



"Fig.1 Proposed Architecture"

The suggested system architecture initiates with the Cresci dataset, succeeded by data preprocessing to cleanse and prepare the inputs. Feature engineering and selection are utilized to derive significant tweet, user, and emotion attributes for representation. The processed data is divided into training, validation, and testing sets, producing feature vectors for modeling purposes. Diverse classifiers with fine-tuning are utilized to differentiate between human and bot accounts. Ultimately, explainable AI methodologies (LIME, SHAP) offer transparency by elucidating model decisions, hence augmenting trust and interpretability in bot identification.

a) Dataset Collection:

i) Cresci-17: The Cresci-17 dataset was compiled to examine social spambots, conventional spambots, counterfeit followers, and authentic accounts on Twitter. The dataset consists of 14,368 users and 6,637,616 tweets, including 10,894 bots and 3,474 humans at the user level, as well as 3,798,254 bot tweets compared to 2,839,362 human tweets. Various account categories, encompassing social spambots, conventional spambots, and counterfeit followers, were incorporated. The profiles and tweet attributes of each user were extracted, aggregated, and labeled, resulting in a comprehensive and balanced dataset appropriate for supervised machine learning and interpretable AI-driven bot identification applications.



Fig.2 Cresci-17 dataset

ii) Cresci-15: The Cresci-15 dataset amalgamates five Twitter datasets—E13, TFP, FSF, INT, and TWT—encompassing 5,301 people and 2,827,757 tweets. It comprises 1,950 human accounts and 3,351 bot accounts. Each user profile comprises 35 raw attributes, including follows, friends, statuses, favorites, listed count, verification status, profile details, and account creation date. Tweets are enhanced with textual attributes like as word count, hashtags, mentions, URLs, emoticons, punctuation, sentiment, and engagement analytics. Aggregated user-level features produce 29 complete variables, facilitating effective bot detection and behavioral analysis.

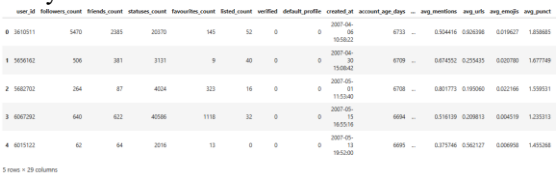


Fig.3 Cresci-15 dataset

b) Pre-Processing:

The subsequent preparation procedures were implemented on both the Cresci-15 and Cresci-17 datasets, methodically preparing, converting, and optimizing user and tweet data to guarantee reliable and precise classification of bot and human accounts.

1. Data Processing: The raw user and tweet data from the Cresci-15 and Cresci-17 datasets were amalgamated across all sub-datasets. Absent numeric values, including follows, friends, statuses, and favorites, were substituted with zeros. The dates of account creation were transformed into datetime format, and the account age in days was computed. Labels were standardized (0=human, 1=bot), guaranteeing uniform identification of accounts across both datasets for subsequent feature engineering and model training.

2. Feature Engineering: User and tweet data were converted to extract significant features. Metrics such as the followers-to-friends ratio, description length, presence of a description, and sentiment scores were calculated for users. Tweet-level attributes encompassed word counts, hashtags, mentions, URLs, emojis, punctuation, unique word ratio, capitalization ratio, tweet duration, and sentiment analysis. Aggregated user statistics, including average hashtags, mentions, emojis, engagement, and sentiment, were computed, yielding a comprehensive feature set for modeling.

3. Visualization: The feature distributions of both datasets were analyzed to discern patterns, ranges, and variability. Histograms, boxplots, and scatter

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

plots were utilized for quantitative attributes including followers, friends, tweet length, and account age. Visual analysis facilitated the identification of outliers, skewed distributions, and distinctions between bots and people, hence informing preprocessing and normalization strategies and enabling astute insights for feature selection and model development.

4. Feature Selection: Selected quantitative and aggregated tweet features were optimized to enhance predictive efficacy. Redundant IDs and columns of minimal informational value were eliminated. Attributes including follower count, friend count, follower-friend ratio, sentiment of descriptions, average hashtags, mentions, emoticons, engagement metrics, and tweet-level analytics were preserved. Feature selection guaranteed that models concentrate on the most informative qualities, hence optimizing efficiency, diminishing noise, and improving classification accuracy for both datasets.

5. Data Balancing: The datasets demonstrated a class imbalance, with a greater number of bots than people. To rectify this, oversampling of the minority class or synthetic data production was utilized, assuring equitable representation of both human and bot accounts. Equitable datasets mitigated model bias towards the predominant class and enhanced prediction efficacy. User-level features were aggregated to ensure consistency and provide equitable training data for precise classification in both Cresci-15 and Cresci-17.

6. Scaling: Numerical attributes including as followers, friends, statuses, favorites, tweet counts, and engagement metrics were standardized using Min-Max or z-score normalization. Standardizing feature ranges mitigated the impact of outliers and ensured that all characteristics contributed equitably to model training. This enhanced convergence for gradient-based classifiers preserved comparability across features, hence improving overall predicted accuracy for both the Cresci-15 and Cresci-17 datasets.

c) Training and Testing:

The processed and feature-engineered data were divided into training and testing sets using stratified sampling to preserve class distribution. User-specific and aggregated tweet attributes were employed to train supervised classifiers. The evaluation of model performance utilized criteria like accuracy, precision, recall, and F1-score. This facilitated strong generalization across the Cresci-15 and Cresci-17 datasets, allowing for dependable identification of bot and human accounts through extensive feature representations.

d) Algorithms:

Decision Tree: Decision Tree evaluates both feature-based and text-based qualities by recursively partitioning data according to the most informative features. It offers a comprehensible framework, facilitating clear classification of accounts while proficiently managing category and numerical factors for the identification of spambots and fraudulent followers.

$$I\,(i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (1)$$

Random Forest: Random Forest generates numerous decision trees and consolidates their results via majority voting, hence enhancing resilience to noise and overfitting. It concurrently assesses several attributes, enhancing generality and precision while effectively detecting automated accounts exhibiting intricate, nonlinear behaviors in social network contexts.

SVM: Support Vector Machine identifies an appropriate hyperplane to enhance class distinction between human and automated accounts. Its efficacy is rooted in high-dimensional analysis utilizing kernel functions, facilitating accurate identification of nuanced feature discrepancies while preserving robustness against overfitting and noisy datasets.

XGBoost: XGBoost constructs sequential boosted decision trees by gradient optimization, effectively capturing complex feature interactions and enhancing accuracy compared to conventional models. It swiftly processes both structured and textual features, scales to extensive datasets, effectively manages imbalance, and guarantees robust performance in spambot detection.

$$\hat{y}_i = \sigma\left(\sum_{k=1}^{K} f_k(x_i)\right), f_k \in F \quad (2)$$

Naïve Bayes: Naïve Bayes employs probability-based categorization by utilizing conditional independence among features. It rapidly analyzes extensive text and structured data, categorizing accounts according to probability. Despite its simplicity, it provides efficiency, functioning as a

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

lightweight and dependable foundation for bot identification tasks.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (3)$$

AdaBoost: AdaBoost integrates sequential weak classifiers, accentuating misclassified instances with adjustable weights to enhance detection efficacy. It enhances prediction consistency, mitigates bias, and adeptly manages imbalance while preserving interpretability, ultimately resulting in a composite model optimally designed for identifying spambots within noisy datasets.

$$H(x) = sign\left(\sum_{t=1}^{T} \propto_t h_t(x)\right) \quad (4)$$

Stacking Classifier: Stacking amalgamates several base models by training a meta-classifier on their outputs. It consolidates many algorithmic strengths, mitigates individual shortcomings, and improves overall detection accuracy by examining both feature-based and text-based attributes, thereby ensuring thorough identification of fraudulent and automated accounts.

$$\hat{y} = g(Y_{base}) = g\left(f_1(x), f_2(x), \ldots, f_m(x)\right) (5)$$

Voting Classifier: Voting amalgamates the results of various classifiers via majority consensus, enhancing decision stability and mitigating errors from individual models. By utilizing complementary algorithmic approaches, it provides balanced and reliable identification of spambots and counterfeit followers while ensuring clear interpretability of outcomes.

$$\hat{y} = argmax_c\left(\sum_{i=1}^{n} II\ (\hat{y}_i = c)\right) \quad (6)$$

e) Integration of XAI and Flask Framework

The amalgamation of Explainable Artificial Intelligence (XAI) with the Flask framework facilitates transparency and accessibility in machine learning applications. XAI methodologies, such SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), offer elucidative insights into model predictions by emphasizing feature contributions and decision-making mechanisms. Incorporating these explanations enables consumers to comprehend the rationale behind a model's classification of an account as either a bot or a human, hence enhancing trust and responsibility.

Flask, a minimalist Python web framework, functions as the interface for deploying these models and visual elucidations. It facilitates smooth interaction between the backend AI models and the frontend user interface. Users can enter data, obtain forecasts, and access comprehensive XAI-generated explanations in real-time, rendering the system intuitive, transparent, and user-centric.

4. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its capacity to appropriately distinguish between patient and healthy cases. To assess the accuracy of a test, one must compute the ratio of true positives and true negatives across all assessed cases. This can be expressed mathematically as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} (7)$$

Precision: Precision assesses the proportion of accurately classified cases among those identified as positive. Consequently, the formula for calculating precision is expressed as:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} (8)$$

Recall: Recall is a metric in machine learning that assesses a model's capacity to recognize all pertinent instances of a specific class. It is the ratio of accurately predicted positive observations to the total actual positives, offering insights into a model's efficacy in identifying occurrences of a specific class.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} (9)$$

F1-Score: The F1 score is a metric for evaluating the accuracy of a machine learning model. It amalgamates the precision and recall metrics of a model. The accuracy metric quantifies the frequency of true predictions generated by a model throughout the entire dataset.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 (10)$$

Table.1 Performance Evaluation – Cresci – 15

ML Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.979	0.979	0.979	0.979
Random Forest	0.993	0.993	0.993	0.993
SVM	0.988	0.988	0.988	0.988

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research

Paper

XGBoost	0.995	0.995	0.995	0.995
Naïve Bayes	0.938	0.943	0.938	0.938
AdaBoost	0.994	0.994	0.994	0.994
Stacking Classifier	0.999	0.999	0.999	0.999
Voting Classifier	0.998	0.998	0.998	0.998

Table 1 presents comparative results indicating that the Stacking Classifier attained the highest overall performance among all assessed methods.

Fig.4 Comparison Graph – Cresci - 15



Figure 4 illustrates the Stacking Classifier as the most effective algorithm, with Accuracy shown in green, Precision in dark green, Recall in light blue, and F1-Score in red across all models.

Table.2 Performance Evaluation – Cresci – 17

ML Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.975	0.975	0.975	0.975
Random Forest	0.990	0.990	0.990	0.990
SVM	0.954	0.955	0.954	0.954
XGBoost	0.991	0.991	0.991	0.991
Naïve Bayes	0.689	0.799	0.689	0.657
AdaBoost	0.987	0.987	0.987	0.987
Stacking Classifier	0.995	0.995	0.995	0.995
Voting Classifier	0.987	0.988	0.987	0.988

Table 2 indicates that the Stacking Classifier attained superior overall performance relative to the other assessed machine learning models.

Fig.5 Comparison Graph – Cresci – 17

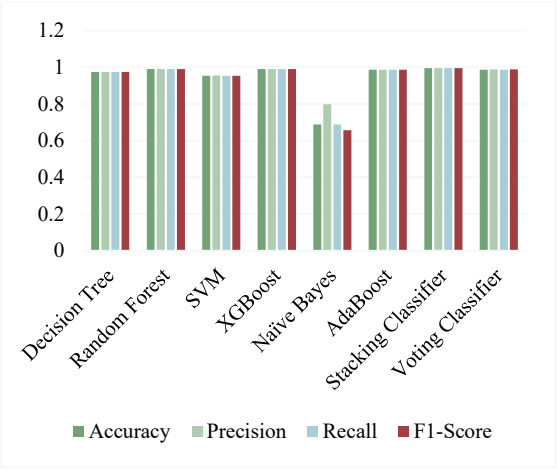


Figure 5 illustrates the Stacking Classifier as the superior algorithm, evaluating eight classifiers based on four metrics: Accuracy (green), Precision (light green), Recall (light blue), and F1-Score (brown).

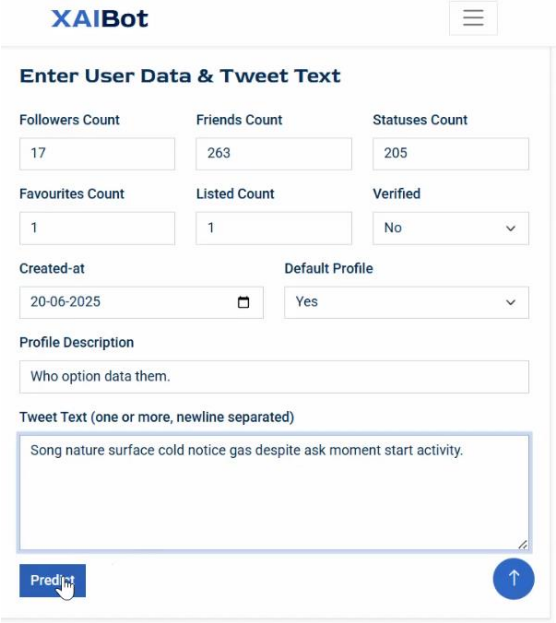


Fig.6 Enter Input Data

Figure 6 illustrates the Twitter interface where users enter profile information and tweet content, encompassing follower counts, friend connections, and status updates.

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper



Fig.7 Predicted Results

Figure 7 illustrates the prediction result, denoting 'Bot' alongside the associated feature values inputted for analysis and classification.



Fig.8 Enter Input Data

Figure 8 illustrates an interface in which users input a username to ascertain whether the account is operated by a bot or a human.

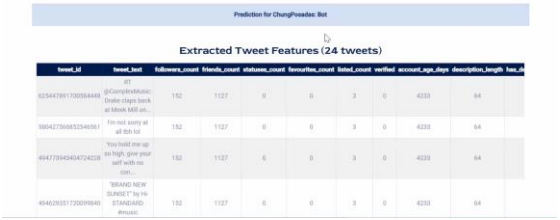


Fig.9 Predicted Results

Figure 9 illustrates the prediction result, indicating 'Bot' for the inputted username alongside pertinent account feature information.

5. CONCLUSION

The research illustrates that the integration of feature-based and text-based data with sophisticated machine learning methodologies may proficiently identify spambots and counterfeit followers on social media platforms. Utilizing Recursive Feature Elimination (RFE) for feature reduction and mitigating class imbalance with SMOTE and SMOTEENN enabled the models to get superior predictive performance. Ensemble learning techniques, especially the Stacking Classifier, achieved optimal outcomes, reaching 99.9% accuracy on the Cresci-15 dataset and 99.5% accuracy on the Cresci-17 dataset, underscoring the efficacy of integrating numerous methods. The incorporation of explainable AI methodologies, such as LIME and SHAP, yielded comprehensive insights into feature significance, guaranteeing interpretability while maintaining high accuracy. The findings validate that sophisticated ensemble models, utilizing meticulously chosen characteristics and equilibrated datasets, can markedly enhance the identification of automated accounts, whereas interpretable methodologies augment comprehension of decision-making processes. The findings indicate that integrating feature selection, sampling strategies, and explainable machine learning techniques enhances detection accuracy while improving transparency and reliability, offering a holistic solution for monitoring and mitigating the effects of spambots and fake followers in social networks.

Future research may concentrate on enhancing the detection framework to accommodate the evolution of more complex automated accounts that perpetually modify their behavior to avoid detection. Integrating real-time data feeds from many social networks might improve scalability and reactivity, enabling models to identify spambot activity as it happens. Advanced deep learning methodologies, such as transformer-based language models, may be investigated to identify intricate textual patterns in conjunction with feature-based signals. Furthermore, using multimodal data, including photos, videos, and network interactions, could enhance detection precision. Continued investigation into explainable AI methodologies will enhance transparency in progressively intricate models, allowing users and administrators to accurately comprehend predictions. Enhancing computational efficiency for implementation in extensive social network settings will be a crucial focus for advancement.

REFERENCES

[1] Alkathiri, N., & Slhoub, K. (2025). Challenges in machine learning-based social bot detection: a systematic review. Discover Artificial Intelligence, 5(1), 214.

[2] Akhtar, M. M., Bhuiyan, N. S., Masood, R., Ikram, M., & Kanhere, S. S. (2025). BotSSCL: Social Bot Detection with Self-Supervised Contrastive Learning. Online Social Networks and Media, 48, 100318.

[3] Lopez-Joya, S., Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2024). Exploring social bots: a feature-based approach to improve bot detection in social networks. arXiv preprint arXiv:2411.06626.

International Journal of
AI Electronics and Nexus Energy
Peer Reviewed, Referred & Indexed Journal
ISSN: 3070-0515 www.zesterapublications.com Original Research
Paper

[4] Nguyen, H. D., Nguyen, D. Q., Nguyen, C. D., To, P. T., Nguyen, D. H., Nguyen-Gia, H., ... & Quan, T. (2024). Supervised learning models for social bot detection: Literature review and benchmark. Expert Systems with Applications, 238, 122217.

[5] Sallah, A., Agoujil, S., Wani, M. A., Hammad, M., Maleh, Y., & Abd El-Latif, A. A. (2024). Fine-tuned understanding: Enhancing social bot detection with transformer-based classification. IEEE Access, 12, 118250-118269.

[6] Mou, G., & Lee, K. (2020, October). Malicious bot detection in online social networks: arming handcrafted features with deep learning. In International conference on social informatics (pp. 220-236). Cham: Springer International Publishing.

[7] G, Viswanath., N, Madhvik., K, Bhaskar., K, Supriya. (2024). Machine-Learning-Based Cloud Intrusion Detection. International Journal of Mechanical Engineering Research and Technology, 16(9), 38–52.

[8] Deshmukh, A., Moh, M., & Moh, T. S. (2024, December). Bot Detection in Social Media Using GraphSage and BERT. In 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 804-811). IEEE.

[9] Pham, P., Nguyen, L. T., Vo, B., & Yun, U. (2022). Bot2Vec: A general approach of intra-community oriented representation learning for bot detection in different types of social networks. Information Systems, 103, 101771.

[10] Javed, D., Jhanjhi, N. Z., & Khan, N. A. (2023, July). Explainable Twitter bot detection model for limited features. In IET Conference Proceedings CP837 (Vol. 2023, No. 11, pp. 476-481). Stevenage, UK: The Institution of Engineering and Technology.

[11] Mbona, I., & Eloff, J. H. (2023). Classifying social media bots as malicious or benign using semi-supervised machine learning. Journal of Cybersecurity, 9(1), tyac015.

[12] Yang, K. C., Varol, O., Hui, P. M., & Menczer, F. (2020, April). Scalable and generalizable social bot detection through data selection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 01, pp. 1096-1103).

[13] Lopez-Joya, S., Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2025). Dissecting a social bot powered by generative AI: anatomy, new trends and challenges. Social Network Analysis and Mining, 15(1), 7.

[14] Pham, P. (2025). An Integrated Pre-Trained Auto-Encoder and Graph Neural Network for General Social Bot Detection. Annals of Data Science, 1-21.

[15] Terumalasetti, S., & Reeja, S. R. (2024). Enhancing social media user's trust: A comprehensive framework for detecting malicious profiles using multi-dimensional analytics. IEEE Access.

[16] Paudel, P., Nguyen, T. T., Hatua, A., & Sung, A. H. (2019, August). How the tables have turned: Studying the new wave of social bots on Twitter using complex network analysis techniques. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 501-508).

[17] Viswanath, G., Abirami, V., & Prathyusha, G. (2024). Hybrid Feature Extraction With Machine Learning To Identify Network Attacks. International Journal Of HRM And Organizational Behavior, 12(3), 217–228.

[18] Wu, J., Ye, X., & Mou, C. (2023). Botshape: A novel social bots detection approach via behavioral patterns. arXiv preprint arXiv:2303.10214.

[19] Bebensee, B., Nazarov, N., & Zhang, B. T. (2021). Leveraging node neighborhoods and egograph topology for better bot detection in social graphs. Social Network Analysis and Mining, 11(1), 10.

[20] Dimitriadis, I., Dialektakis, G., & Vakali, A. (2024). Caleb: a conditional adversarial learning framework to enhance bot detection. Data & Knowledge Engineering, 149, 102245.

[21] Guo, S., Wang, J., Wang, Z., Yu, G., & Wu, S. (2025). BotICC: enhancing social bot detection through implicit connection computation. Multimedia Systems, 31(3), 210.

[22] Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

[23] Moghaddam, S. H., & Abbaspour, M. (2022). Friendship preference: Scalable and robust category of features for social bot detection. IEEE Transactions on Dependable and Secure Computing, 20(2), 1516-1528.

[24] Ellaky, Z., Benabbou, F., Ouahabi, S., & Sael, N. (2021, June). A survey of spam bots detection in

online social networks. In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA) (pp. 58-65). IEEE.

[25] Lakshmi, J. M., Prasad, K. K., & Viswanath, G. (2025). Proactive Security in Multi-Cloud Environments: A Blockchain Integrated Real-Time Anomaly Detection and Mitigation Framework. Cuestiones De Fisioterapia, 54(2), 392–417.

[26] Najari, S., Salehi, M., & Farahbakhsh, R. (2022). GANBOT: a GAN-based framework for social bot detection. Social Network Analysis and Mining, 12(1), 4.

[27] G Loge, T Sunil Kumar Reddy, G Swapna, & G Viswanath. (2025). Interpretable AI for Precision Brain Tumor Prognosis: A Transparent Machine Learning Approach. In International Journal of Health Sciences and Pharmacy (IJHSP) (Vol. 9, Number 1, pp. 180–195). Zenodo. https://doi.org/10.5281/zenodo.15523628

[28] Kouvela, M., Dimitriadis, I., & Vakali, A. (2020, November). Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities. In Proceedings of the 12th International Conference on Management of Digital EcoSystems (pp. 55-63).

[29] Li, Y., Li, Z., Gong, D., Hu, Q., & Lu, H. (2024). BotCL: a social bot detection model based on graph contrastive learning. Knowledge and Information Systems, 66(9), 5185-5202.

[30] Kumar, A. S., Kumar, N. S., Devi, R. K., & Muthukannan, M. (2023). Analysis of deep learning-based approaches for spam bots and cyberbullying detection in online social networks. AI-Centric Modeling and Analytics, 324-361.