Identification of Spambots and Fake Followers on Social Network via Interpretable AI-Based Machine Learning

Mohammed Abdul Hai Zubair
PG Scholar, Department of Computer Science and Engineering,
Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086
zubair4u1@gmail.com

Subramanian K.M
'Professor, Department of Computer Science and Engineering,
Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086
kmsubbu.phd@gmail.com

Md. Ateeq Ur Rahman
Professor, Department of Computer Science and Engineering:
Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086
mail_to_ateeq@yahoo.com

***Abstract*— Social networking platforms such as Twitter enable broad human interaction but are increasingly targeted by automated accounts that mimic human behavior, spreading misinformation and manipulating public opinion. Detecting such spambots is critical for maintaining information integrity, yet many conventional approaches rely on opaque, black-box models, limiting transparency and interpretability. This study employs the Cresci-15 and Cresci-17 datasets to investigate interpretable machine learning techniques for identifying spambots and fake followers. Both feature-based and text-based data are utilized, with preprocessing steps including normalization, tokenization, and removal of irrelevant content. Recursive Feature Elimination (RFE) reduces feature dimensionality, while resampling strategies such as SMOTE and SMOTEENN address class imbalance. Multiple machine learning algorithms, including Decision Tree, Random Forest, SVM, XGBoost, AdaBoost, Stacking Classifier, and Voting Classifier, are evaluated. Results demonstrate that Stacking Classifier achieves superior performance, reaching 99.9% accuracy on Cresci-15 and 99.5% accuracy on Cresci-17 datasets. Additionally, explainable AI methods such as LIME and SHAP provide clear insights into feature importance, enhancing model transparency and supporting informed decision-making. These findings highlight the effectiveness of combining feature selection, advanced resampling, and ensemble learning strategies with interpretable techniques for robust detection of automated accounts in social networks.**

Keywords*— *Interpretable AI, social network, bot detection, fake followers, spambots."

I. INTRODUCTION

Social networks have emerged as the central hub of information dissemination in the modern digital era. Among them, X (formerly Twitter) has become one of the most prevalent and influential platforms, enabling millions of active users to connect and engage in real-time conversations [1]. Its vast social and economic impact, however, has also attracted malicious actors who exploit its open nature to manipulate public opinion and spread disinformation. Automated programs, known as bots, are among the most common tools used for this purpose. While some bots are beneficial, generating useful content such as educational blogs and news updates, malicious bots are designed to disseminate spam, rumors, and harmful material [2].

The detection of these malicious accounts has become a crucial challenge in ensuring the integrity of online discourse. Bot detection strategies often rely on user metadata, behavioral traits, timestamps, and network structures [3]. Yet, the process of feature engineering—designing and extracting these characteristics—demands extensive effort and domain expertise. Moreover, bots continue to evolve, adopting adaptive strategies that mimic human behavior, making them increasingly difficult to detect. Malicious bots also facilitate the rapid amplification of misinformation, including fake news and hate speech, by strategically interacting with high-profile accounts [4].

The ecosystem of automated manipulation is further complicated by botnets and Sybil accounts. A botnet refers to a network of coordinated bots performing specific malicious tasks, while Sybil accounts represent fabricated identities used to distort authentic conversations [5], [6]. Such practices disrupt genuine engagement and amplify low-credibility content, raising serious concerns regarding trust and reliability in social media ecosystems.

Given these threats, machine learning (ML) has been widely employed in domains such as sports analytics [7], sentiment analysis [8], [9], and fake news detection [10]. In recent years, ML-based approaches have been extended to social bot detection, offering automated solutions that outperform traditional heuristic and network-based methods. However, many ML-based detection frameworks still operate as "black-box" models, providing limited interpretability. This lack of transparency makes it challenging for researchers and practitioners to understand whether predictions are based on meaningful patterns or overfitted noise.

To address these limitations, interpretable machine learning (XAI) is increasingly being integrated into bot detection systems. XAI enhances transparency by explaining the contribution of individual features to model predictions, thereby improving accountability and trust. By bridging the gap between algorithmic decisions and human comprehension, XAI-based frameworks provide both accuracy and interpretability, paving the way for more robust and trustworthy detection of bots on X and similar social networks.

II. RELATED WORK

The detection of bots in social networks has become a critical research area due to the evolving nature of malicious accounts that manipulate online discourse and information ecosystems. Researchers have increasingly explored advanced machine learning, graph-based methods, and interpretable frameworks to improve detection accuracy and scalability.

Mbona and Eloff [11] highlighted the need to classify bots not only as malicious but also as benign, recognizing that many bots serve neutral or positive purposes such as news delivery. They proposed a semi-supervised learning approach that leverages limited labeled data with abundant unlabeled data, reducing dependence on costly annotations while enhancing differentiation between harmful and benign bots. Similarly, Yang et al. [12] addressed scalability and generalizability by introducing a data selection strategy that ensures machine learning models trained on representative datasets perform effectively across multiple contexts, a crucial step toward building adaptable detection systems.

The rise of generative AI-driven bots introduces new challenges. Lopez-Joya et al. [13] analyzed the anatomy of bots powered by large language models, showing that their ability to generate realistic, contextually coherent text makes them harder to detect than traditional spambots. They emphasized the need for updated strategies that can adapt to AI-enhanced adversaries. Complementing this, Pham [14] proposed an integrated model combining a pre-trained auto-encoder with graph neural networks (GNNs), which captures both latent behavioral features and structural social interactions. This hybrid approach improved generalizability across networks, outperforming conventional supervised techniques.

Terumalasetti and Reeja [15] emphasized enhancing user trust by developing a multi-dimensional analytics framework that integrates content, temporal, and relational features. Their holistic design provided robustness against adaptive spambots, reflecting the necessity of multidimensional models rather than reliance on single feature categories. Earlier, Paudel et al. [16] had demonstrated how bot strategies evolved to mimic human-like interactions within networks. Using complex network analysis, they showed that structural metrics like clustering coefficients

and centrality measures are valuable indicators for detecting sophisticated bots.

Aljabri et al. [17] synthesized progress in the field through a comprehensive literature review of machine learning-based bot detection. They highlighted challenges including dependence on feature engineering, scalability limitations, and the opacity of black-box models. Their review underscored the importance of balancing detection accuracy with interpretability and adaptability. Wu et al. [18] extended this perspective with Botshape, a framework emphasizing behavioral pattern analysis such as posting frequencies and temporal activity styles. By modeling dynamic behaviors, their approach achieved strong performance against bots designed to mimic human timing.

Graph-based approaches have also gained traction. Bebensee et al. [19] proposed leveraging node neighborhoods and egograph topology to capture relational differences at the micro-structural level of social graphs, reducing reliance on handcrafted features. Their work showed the potential of graph-based learning to detect anomalies within contextualized network structures. Finally, Dimitriadis et al. [20] introduced Caleb, a conditional adversarial learning framework that improves robustness against adaptive bot strategies. By training on adversarial examples, their method enhanced generalization and resilience, directly addressing the evolving arms race between detection systems and adversaries.

III. Materials And Methods

The proposed system introduces a robust and interpretable machine learning framework for detecting spambots and fake followers on social networks by integrating both feature-based and text-based data from the Cresci-15 and Cresci-17 datasets. Data preprocessing includes handling missing values, encoding categorical variables, standardizing numerical features, cleaning text (removal of HTML tags, URLs, non-printable characters), normalization, tokenization, and TF-IDF representation. To address class imbalance, SMOTE and SMOTEENN are applied. Predictive models are built using Decision Tree, Random Forest, SVM, XGBoost, Naïve Bayes, and AdaBoost, enhanced with Stacking and Voting Classifiers. Recursive Feature Elimination reduces dimensionality, while LIME and SHAP ensure interpretability. A Flask-based interface enables real-time detection and interactive analysis of model decisions.

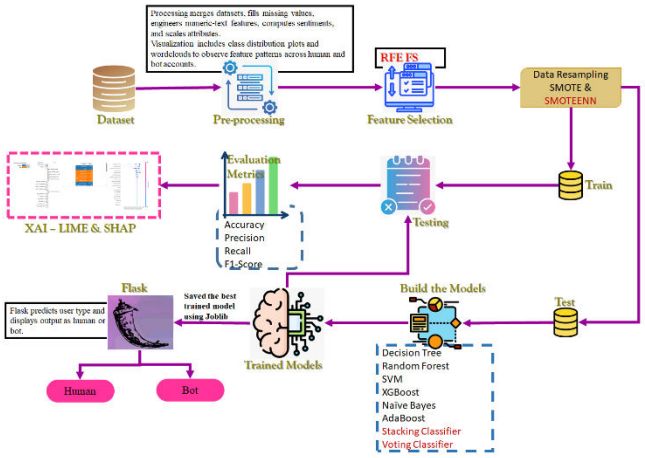

Fig. 1. System Architecture

The proposed system architecture begins with the Cresci dataset, followed by data preprocessing to clean and prepare inputs. Feature engineering and selection are applied to extract meaningful tweet, user, and sentiment features for representation. The processed data is split into training, validation, and testing sets, generating feature vectors for modeling. Various classifiers with fine-tuning are employed to distinguish between human and bot accounts. Finally, explainable AI techniques (LIME, SHAP) provide transparency by explaining model decisions, enhancing trust and interpretability in bot detection.

a) Dataset Collection:

i) Cresci-17: The Cresci-17 dataset was collected to study social spambots, traditional spambots, fake followers, and genuine accounts on Twitter. It comprises 14,368 users and 6,637,616 tweets, with 10,894 bots and 3,474 humans at the user level, and 3,798,254 bot tweets versus 2,839,362 human tweets. Multiple account types, including social spambots, traditional spambots, and fake followers, were included. Each user's profile and tweet features were extracted, aggregated, and labeled, providing a comprehensive, balanced dataset suitable for supervised machine learning and interpretable AI-based bot detection tasks.

	user_id	followers_count	friends_count	statuses_count	favourites_count	listed_count	verified	default_profile	created_at	account_age_days	description	description_len
0	1502026416	208	332	2177	265	1	0	0	2013-06-11 11:20:35	4474	15years ago X.Lines24	
1	2492782375	330	485	2660	3972	5	0	1	2014-05-13 10:37:57	4138	[illegible]	
2	293212315	166	177	1254	1185	0	0	0	2011-05-04 23:30:37	5243	Let me see what your best move is!	
3	191839658	2248	981	202968	60304	101	0	0	2010-09-17 14:02:10	5472	20. menna: #farida #nyc and the 80s actually y...	
4	3020965143	21	79	82	5	0	0	1	2015-02-06 04:10:49	3870	Cosmetologist	

Fig.2 Cresci-17 dataset

ii) Cresci-15: The Cresci-15 dataset consolidates five Twitter datasets—E13, TFP, FSF, INT, and TWT—comprising 5,301 users and 2,827,757 tweets. It includes 1,950 human and 3,351 bot accounts. Each user profile contains 35 raw features such as followers, friends, statuses, favourites, listed count, verified status, profile details, and account creation date. Tweets are enriched with text-based features like word count, hashtags, mentions, URLs, emojis, punctuation, sentiment, and engagement metrics. Aggregated user-level features yield 29 comprehensive variables, enabling robust bot detection and behavioral analysis.

	user_id	followers_count	friends_count	statuses_count	favourites_count	listed_count	verified	default_profile	created_at	account_age_days	description	description_length
0	3610511	5470	2385	20370	145	52	0	0	2007-04-06 10:58:22	6733	Founder of http://www.screenweek.it & http://w...	151
1	5656162	506	381	3131	9	40	0	0	2007-04-30 15:08:42	6709	BSc degree (cum laude) in Computer Engineering...	104
2	5682702	264	87	4024	323	16	0	0	2007-05-01 11:53:40	6708	Cogito ergo bestemmio.	22
3	6067292	640	622	40586	1118	32	0	0	2007-05-15 16:55:16	6694	Se la vita ti dÃ sarde, scapocciale!	37
4	6015122	62	64	2016	13	0	0	0	2007-05-13 19:52:00	6695	Je me souviens	14

Fig.3 Cresci-15 dataset

b) Pre-Processing:

The following preprocessing steps were applied to both Cresci-15 and Cresci-17 datasets, systematically preparing, transforming, and optimizing user and tweet data to ensure robust, accurate bot and human account classification.

1. Data Processing: Raw user and tweet data from Cresci-15 and Cresci-17 datasets were combined across all sub-datasets. Missing numeric values such as followers, friends, statuses, and favorites were filled with zeros. Account creation dates were converted to datetime format, and account age in days was calculated. Labels were standardized (0=human, 1=bot), ensuring consistent identification of accounts across both datasets for downstream feature engineering and model training.

2. Feature Engineering: User and tweet data were transformed to derive meaningful features. For users, metrics like followers-to-friends ratio, description length, presence of description, and sentiment scores were computed. Tweet-level features included word counts, hashtags, mentions, URLs, emojis, punctuation, unique word ratio, capitalization ratio, tweet length, and sentiment. Aggregated statistics per user, such as average hashtags, mentions, emojis, engagement, and sentiment, were calculated, resulting in a comprehensive feature set for modeling.

3. Visualization: Feature distributions for both datasets were inspected to understand patterns, ranges, and variability. Histograms, boxplots, and scatter plots were applied to numeric features such as followers, friends, tweet length, and account age. Visual analysis helped identify outliers, skewed distributions, and differences between bots and humans, guiding preprocessing and normalization decisions and enabling informed insights for feature selection and model development.

4. Feature Selection: Relevant numeric and aggregated tweet features were selected to maximize predictive power. Redundant identifiers and low-informative columns were removed. Features such as followers count, friends count, followers-friends ratio, description sentiment, average hashtags, mentions, emojis, engagement metrics, and tweet-level statistics were retained. Feature selection ensured that models focus on the most informative attributes, improving efficiency, reducing noise, and enhancing classification accuracy for both datasets.

5. Data Balancing: The datasets exhibited class imbalance, with more bots than humans. To address this, oversampling of the minority class or synthetic data generation was applied, ensuring balanced representation of both human and bot accounts. Balanced datasets prevented model bias toward the majority class and improved predictive performance. Aggregated user-level features were used to maintain consistency while providing equitable training data for accurate classification across both Cresci-15 and Cresci-17.

6. Scaling: Numerical features including followers, friends, statuses, favorites, tweet counts, and engagement metrics were scaled using Min-Max or z-score normalization. Scaling standardized feature ranges, minimized the influence of extreme values, and ensured all features contributed proportionally to model training. This improved convergence for gradient-based classifiers and maintained comparability across features, enhancing overall predictive accuracy for both Cresci-15 and Cresci-17 datasets.

c) Training and Testing:

Processed and feature-engineered data were split into training and testing sets with stratified sampling to maintain class distribution. User-level and aggregated tweet features were used to train supervised classifiers. Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. This ensured robust generalization across both Cresci-15 and Cresci-17 datasets, enabling reliable detection of bot and human accounts based on comprehensive feature representations.

C) Algorithms

Decision Tree: Decision Tree analyzes both feature-based and text-based attributes by recursively splitting data on the most informative features. It provides an interpretable structure, enabling transparent classification of accounts while effectively handling categorical and numerical variables for detecting spambots and fake followers.

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

Random Forest: Random Forest constructs multiple decision trees and aggregates their outcomes through majority voting, ensuring robustness against noise and overfitting. It evaluates diverse features simultaneously, improving generalization and accuracy while reliably identifying automated accounts with complex, nonlinear behaviors within social network environments.

SVM: Support Vector Machine determines an optimal hyperplane to maximize class separation between human and automated accounts. Its strength lies in high-dimensional analysis with kernel functions, enabling precise detection of subtle feature differences while maintaining resilience against overfitting and noisy datasets.

XGBoost: XGBoost builds sequential boosted decision trees using gradient optimization, capturing intricate feature interactions and improving accuracy over traditional models. It processes both structured and text-based attributes efficiently, scales to large datasets, handles imbalance effectively, and ensures strong performance for spambot identification.

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

Naïve Bayes: Naïve Bayes applies probability-based classification by leveraging conditional independence among features. It quickly processes large-scale text and structured data, assigning accounts to categories based on likelihood. Despite simplicity, it offers efficiency, serving as a lightweight, reliable baseline for bot identification tasks.

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

AdaBoost: AdaBoost combines sequential weak classifiers, emphasizing misclassified examples with adaptive weights to boost detection strength. It improves prediction consistency, reduces bias, and handles imbalance effectively while maintaining interpretability, ultimately creating a composite model well-suited for detecting spambots across noisy datasets.

Stacking Classifier: Stacking integrates multiple base models by training a meta-classifier on their predictions. It captures diverse algorithmic strengths, reduces individual weaknesses, and enhances overall detection accuracy by analyzing both feature-based and text-based characteristics, ensuring comprehensive identification of fake and automated accounts.

$$\hat{y} = g(Y_{base}) = g\big(f_1(x), f_2(x), \dots, f_m(x)\big) \quad (4)$$

Voting Classifier: Voting combines outputs of multiple classifiers through majority consensus, improving decision stability and reducing errors from individual models. By leveraging complementary algorithmic perspectives, it delivers balanced, reliable detection of spambots and fake followers while maintaining straightforward interpretability of results.

$$\hat{y} = argmax_c\left(\sum_{i=1}^{n} II\ (\hat{y}_i = c)\right) \quad (5)$$

e) Integration of XAI and Flask Framework:

The integration of Explainable Artificial Intelligence (XAI) with the Flask framework enables both transparency and accessibility in machine learning applications. XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), provide interpretable insights into model predictions, highlighting feature contributions and decision-making processes. By incorporating these explanations, users can understand why a model classifies an account as a bot or human, increasing trust and accountability.

Flask, a lightweight Python web framework, serves as the interface for deploying these models and visual explanations. It allows seamless interaction between the backend AI models and the front-end user interface. Users can input data, receive predictions, and view detailed XAI-generated explanations in real-time, making the system intuitive, transparent, and user-friendly.

IV. Experimental Results

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (6)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (7)$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 \quad (9)$$

Table.1 Performance Evaluation – Cresci – 15

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

In Table.1, comparative results highlight that the Stacking Classifier achieved the highest overall performance among all evaluated algorithms.

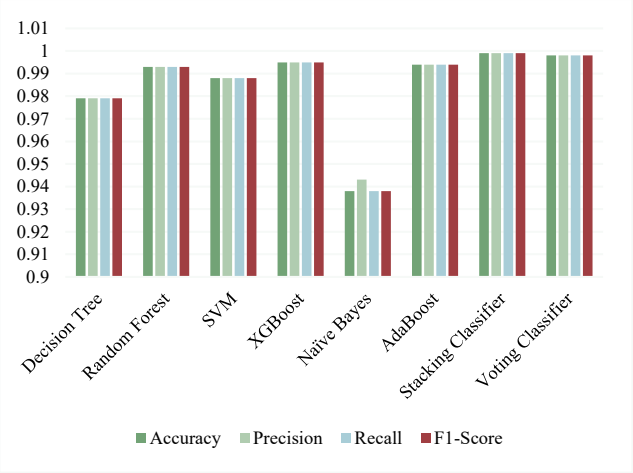

Fig.4 Comparison Graph – Cresci – 15

In fig.4 highlights the Stacking Classifier as the highest-performing algorithm, where Accuracy is green, Precision is dark green, Recall is light blue, and F1-Score is 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
RF	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

In Table.2, results show that the Stacking Classifier achieved the highest overall performance compared to other evaluated machine learning models.

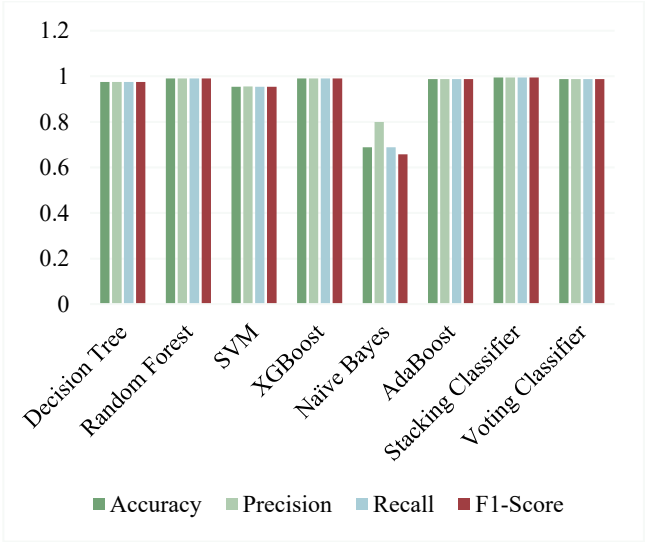

Fig.5 Comparison Graph – Cresci – 17

Fig. 5 highlights the Stacking Classifier as the top-performing algorithm, comparing eight classifiers across four metrics: Accuracy (green), Precision (light green), Recall (light blue), and F1-Score (brown).

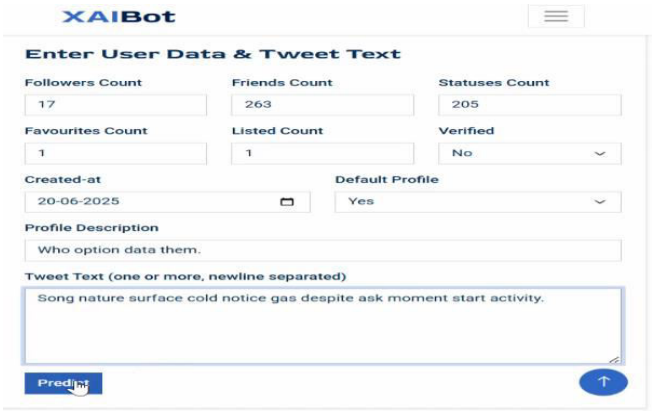

Fig.6 Enter Input Data

Fig. 6 shows the Twitter interface where users input profile details and tweet text, including followers, friends, and status counts.

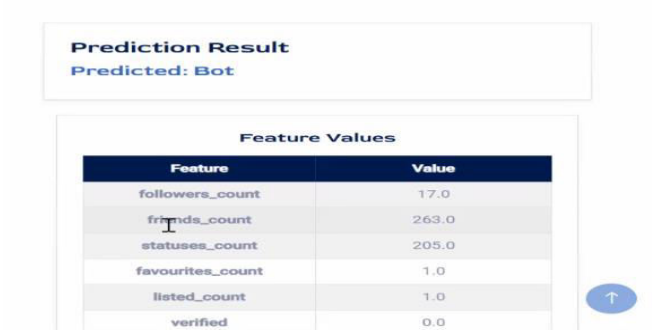

Fig.7 Predicted Results

Fig. 7 displays the prediction output, indicating 'Bot' along with corresponding feature values entered for analysis and classification.

Fig.8 Enter Input Data

Fig. 8 shows an interface where users enter a username to determine whether the account is a bot or a human.

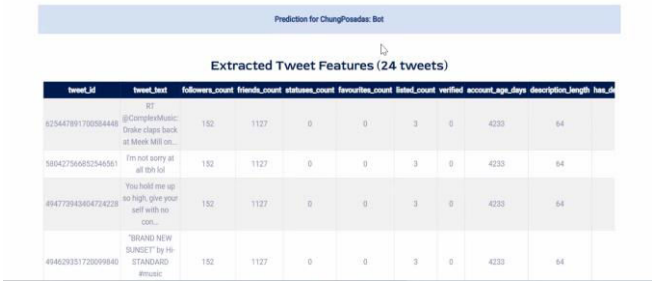

Fig.9 Predicted Results

Fig.9 displays the prediction outcome, showing 'Bot' for the entered username along with relevant account feature details.

V. Conclusion

The study demonstrates that combining feature-based and text-based data with advanced machine learning techniques can effectively detect spambots and fake followers on social networks. By employing Recursive Feature Elimination (RFE) for feature reduction and addressing class imbalance through SMOTE and SMOTEENN, the models were able to achieve high predictive performance. Ensemble learning strategies, particularly the Stacking Classifier, delivered the best results, attaining 99.9% accuracy on the Cresci-15 dataset and 99.5% accuracy on the Cresci-17 dataset, highlighting the robustness of combining multiple algorithms. Integration of explainable AI techniques, including LIME and SHAP, provided detailed insights into feature importance, ensuring interpretability alongside high accuracy. The results confirm that advanced ensemble models with carefully selected features and balanced datasets can significantly improve the detection of automated accounts, while interpretable methods enhance understanding of decision-making processes. Overall, the findings suggest that incorporating feature selection, sampling strategies, and explainable machine learning approaches not only maximizes detection accuracy but also strengthens transparency and reliability, providing a comprehensive solution for monitoring and mitigating the impact of spambots and fake followers in social network environments.

Future Scope

Future research can focus on extending the detection framework to handle evolving and more sophisticated automated accounts that continuously adapt their behavior to evade identification. Incorporating real-time data streams from multiple social networks can enhance scalability and responsiveness, allowing models to detect spambot activity as it occurs. Advanced deep learning techniques, including transformer-based language models, could be explored to capture complex textual patterns alongside feature-based signals. Additionally, integrating multimodal data, such as images, videos, and network interactions, may improve detection accuracy. Further exploration of explainable AI approaches will help ensure transparency in increasingly complex models, enabling users and administrators to interpret predictions reliably. Optimizing computational efficiency for deployment in large-scale social network environments will also be a key area for improvement.

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] Javed, D., Jhanjhi, N. Z., Khan, N. A., Ray, S. K., Al-Dhaqm, A., & Kebande, V. R. (2025). Identification of spambots and fake followers on social network via interpretable ai-based machine learning. IEEE Access.

[4] Dhanaraj, C., & Shankar, B. (2026). Identification Of Spam Bots And Fake Followers On Social Network Via Interpretable Ai Based Ml. Research Digest on Engineering Management and Social Innovations, 2(5), 486-492.

[5] Dhanaraj, C., & Shankar, B. (2026). Identification Of Spam Bots And Fake Followers On Social Network Via Interpretable Ai Based Ml. Research Digest on Engineering Management and Social Innovations, 2(5), 486-492.

[6] NIDAMANURU, D. S. R., SATHVIKA, A., RAMESH, G. S., REDDY, B. N., HUZAIFA, A., & KUMAR, B. S. (2026). AI-POWERED DETECTION OF FAKE FOLLOWERS AND SOCIAL MEDIA SPAMBOTS USING EXPLAINABLE MACHINE LEARNING. International Journal of Data Science and IoT Management System, 5(1), 753-761.

[7] Palani Murugan, S., Gobinath, R., Rahapriya, K., Fahumitha Afrose, B., Fathima Fazlina, M., & Durga Devi, S. InstaGuard: An AI-Powered Framework for Detecting Fraudulent Activities on Instagram Using Machine Learning and Deep Learning Techniques. Proceedings Copyright, 840, 847.

[8] Selvam, L., Ramya, S., Balamurugan, S., Raj, J. R. F., Fernando, A. V., & Lavanya, S. P. (2026, January). Scalable and Interpretable Engineering Solutions for Fake Account Detection using GNNs and FactorVAE. In 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) (pp. 1127-1135). IEEE.

[9] Mounika, K., & Reddy, N. R. (2025). An integrated machine learning framework for spammer and fake user detection in online social networks. Fringe Multi-Engineering Proceedings (FMEP, ISSN: 3107-6149), 1(3), 12-25.

[10] Prabha[1], M. L., Balaji, K., Madhan, B., & Thulasiram, S. (2026, March). AI Based Fake Challan Detection From Sms Using Machine Learning. In Proceedings of the International Conference on Artificial Intelligence and Secure Data Analytics (ICAISDA 2025) (p. 40). Springer Nature.

[11] Awan, S. F., Wasim, M., Safdar, S., Rehman, A., & Ghani, M. U. (2025, April). A Voting Ensemble Model and Explainability Framework for Fake Account Detection in Social Media. In 2025 International Conference on Emerging Technologies in Electronics, Computing, and Communication (ICETECC) (pp. 1-6). IEEE.

[12] Baazeem, R. (2026). Multi‐Modal Deep Learning and Heterogeneous Graph Reasoning Framework for Robust Social Media Profile Authenticity Verification. Concurrency and Computation: Practice and Experience, 38(11), e70748.

[13] Selvam, L., Ramya, S., Balamurugan, S., Raj, J. R. F., Fernando, A. V., & Lavanya, S. P. (2026, January). Scalable and Interpretable Engineering Solutions for Fake Account Detection using GNNs and FactorVAE. In 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) (pp. 1127-1135). IEEE.

[14] Zardi, H., & Alreshoodi, R. (2026). Cost-Sensitive Fake Profile Detection in Online Social Networks Using Random Forest Feature Selection and LightGBM. Engineering, Technology & Applied Science Research, 16(1), 30906-30912.

[15] Selvam, L., Ramya, S., Balamurugan, S., Raj, J. R. F., Fernando, A. V., & Lavanya, S. P. (2026, January). Scalable and Interpretable Engineering Solutions for Fake Account Detection using GNNs and FactorVAE. In 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) (pp. 1127-1135). IEEE.

[16] Rostok, V., Venables, A., & Nõmm, S. (2026). Machine Learning Approaches for Event Detection and Credibility Analysis in the Digital Information Environment: A Literature Review. IEEE Transactions on Computational Social Systems.

[17] AlZeyadi, H., Sert, R., & Duran, F. (2025). A Lightweight, Explainable Spam Detection System with Rüppell's Fox Optimizer for the Social Media Network X. Electronics, 14(21), 4153.

[18] Khan, A., Fatima, A., Jamil, R., Ahmed, H., & Saba, A. (2026). Enhancing Social Media Bot Detection with Cross-Feature Gating and Residual Learning. ICCK Transactions on Emerging Topics in Artificial Intelligence, 3(1), 20-32.

[19] ALZEYADI, H., SERT, R., & Duran, F. (2025). A Lightweight, Explainable Spam Detection System with Rüppell's Fox Optimizer for the Social Network X.

[20] Khedkar, R. A., Singh, D. M., Gosavi, C., & Hajare, A. (2026, January). Anomaly Detection Methods in Social Networks. In 2026 International Conference on Emerging Trends and Innovations in ICT (ICEI) (pp. 1-6). IEEE.

[21] Prabha[1], M. L., Balaji, K., Madhan, B., & Thulasiram, S. (2026, March). AI Based Fake Challan Detection From Sms Using Machine Learning. In Proceedings of the International Conference on Artificial Intelligence and Secure Data Analytics (ICAISDA 2025) (p. 40). Springer Nature.

[22] Varshney, D., & Grover, M. (2026). Artificial Intelligence in Marketing: Concepts, Strategies and Applications. Chyren Publication.

[23] Shukla, P. K., Veerasamy, B. D., Alduaiji, N., Addula, S. R., Sharma, S., & Shukla, P. K. (2025). Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection. Peer-to-Peer Networking and Applications, 18(4), 232.

[24] Ali, A. S. Artificial Intelligence, Media Technology, and Islamic Identity: Representing the SOAS Mosque as a National and Spiritual Symbol.

[25] Rohith, R., Kaushik, S., Varadharaj, S., Pulivarthi, L., Gulati, P. K., & Uday, A. (2025, November). Advances in Social Media Analytics: A Comprehensive Review of Emerging Computational Approaches. In 2025 Fourth International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN) (pp. 1-8). IEEE.

[26] Ali, A. S. Artificial Intelligence in Documentary Media: Educating Society on Zakat as an Instrument of Social Justice.

[27] Zardi, H., & Alreshoodi, R. (2026). Cost-Sensitive Fake Profile Detection in Online Social Networks Using Random Forest Feature Selection and LightGBM. Engineering, Technology & Applied Science Research, 16(1), 30906-30912.

[28] Malik, B. Artificial Intelligence, Data Privacy, and Digital Media Ethics: Protecting User Rights in the Age of Algorithmic Surveillance.

[29] Tiwari, A., Sinha, A., Verma, S., & Pandey, S. K. (2026, March). The Evolution of Social Bot Detection: A Comprehensive Survey of Methods, Datasets, and Emerging Challenges. In 2026 IEEE International Conference on AI Engineering and Innovations (AIEI) (pp. 1-7). IEEE.

[30] Selvam, L., Ramya, S., Balamurugan, S., Raj, J. R. F., Fernando, A. V., & Lavanya, S. P. (2026, January). Scalable and Interpretable Engineering Solutions for Fake Account Detection using GNNs and FactorVAE. In 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC) (pp. 1127-1135). IEEE.