ETDT

Emerging Trends in Digital Transformation

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-30***

EXPLAINABLE MACHINE LEARNING MODELS FOR SPAMBOT AND FAKE FOLLOWER DETECTION

[#1]**Koni Alekhya,** *Department of MCA,*
[#2]**Mr. S. Sateesh Reddy,** *Associate Professor, Department of CSE,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: This research presents an explainable ML approach to the detection of spambots and false followers across multiple social media platforms by combining model interpretability with classification accuracy. In the investigation, suspicious accounts are identified by analyzing the following: posting frequency, follower-to-follower ratio, engagement consistency, account formation patterns, content similarity, and behavioral and profile-based network interaction characteristics. In order to provide a clear explanation for the predictions' outcomes, the use of advanced machine learning techniques such as Random Forest, XGBoost, and Support Vector Machine, as well as Explainable AI (XAI) methodologies like SHAP and LIME, is employed. The proposed paradigm enhances accountability and confidence by revealing the factors that administrators and analysts employ to make bot detection judgments. The explainable architecture enhances the reliability of automated moderation systems by reducing false positives and increasing detection accuracy, as demonstrated by the experimental results. The research promotes genuine user engagement in online activities, curbs the dissemination of false information, and ensures the security of social media platforms through the implementation of plain and efficient spam detection methods.

Keywords: *Explainable Machine Learning, Spambot Detection, Fake Follower Identification, Social Media Security, Explainable AI (XAI), Random Forest, XGBoost, SHAP,*

1. INTRODUCTION

Modern communication, marketing, and information transmission are now reliant on Facebook, Instagram, and X (formerly known as Twitter). The rapid expansion of these platforms has resulted in the proliferation of spambots and phony follower accounts, which has distorted online influence, disseminated false information, and altered engagement metrics. These accounts are difficult to identify because they often resemble genuine user behavior. Conventional detection systems are incapable of adapting to novel patterns, necessitating the implementation of reliable and intelligent detection methods.

The increasing prevalence of automated accounts being employed for the purpose of spam dissemination, false interaction, and opinion manipulation serves as the impetus for this research. Spambots are notorious for disseminating misleading or harmful content on a large scale, while bogus followers are frequently employed to fraudulently inflate popularity. Machine learning models have improved their detection accuracy; however, many of them remain opaque "black-box" systems that fail to disclose their decision-making processes. This secrecy undermines credibility and impedes the pervasive application of the technology in mission-critical environments.

DOI: https://doi.org/10.5281/zenodo.20603900

This research is most effective when it enhances the accuracy and interpretability of detection. In addition to precise predictions, cybersecurity teams, advertisers, and social media firms seek specific reasons to classify an account as either human or machine. Explainable machine learning endeavors to render model conclusions more transparent and comprehensible in order to address this knowledge divide. This ensures improved regulatory compliance on digital platforms, promotes moral decision-making, and enhances confidence.

Accurate algorithm identification is crucial from a business standpoint to ensure equitable engagement metrics, safeguard user confidence, and preserve platform trust. Fake accounts distort the analytics and advertising results that are based on social media data, resulting in businesses making poor judgments. Although advanced AI models, such as deep learning and ensemble approaches, are more effective at detecting, they are not always straightforward to comprehend. In order to be practical in real-world applications, AI must be explicable.

2. LITERATURE REVIEW

Anderson & Carter (2021): This paper introduces a framework for interpretable machine learning that is capable of detecting spam bots and false followers on social media. Random Forest and Decision Tree classifiers are employed to analyze user activity patterns, follower relationships, and interaction irregularities. AI technologies that are able to provide explanations in common English are beneficial for shaky account predictions. Feature importance analysis can be employed to enhance comprehension of bot behavior and automated spam dissemination. The results of the experiments indicate that the accuracy of detection has improved and the rate of false-positive identification has decreased.

Singh & Walker (2022): This work aims to identify malicious social media bots and false followers by introducing an ensemble learning architecture that is comprehensible. Gradient Boosting and Support Vector Machine are implemented to evaluate the reliability of interactions, the regularity of posts, and the authenticity of profiles. In order to provide real-time explanations for categorization determinations, the framework implements interpretability algorithms that are based on SHAP. By contrasting the results, we have determined that the distinction between genuine users and fabricated accounts is now more precise and dependable. The proposed paradigm can facilitate the establishment of social networking systems that are both secure and reliable.

Garcia & Bennett (2023): This article introduces a hybrid interpretable AI model for the detection of spambots on social networks that employs behavioral analytics and graph-based learning. The method combines network centrality metrics with machine learning classifiers to detect coordinated fraudulent follower activity. The utilization of LIME explanation methodologies contributes to the transparency of prediction outcomes and the increased confidence of users in them. Methods for data purification and anomaly detection are implemented to improve the dependability of extensive datasets. The results of the testing indicate that it is capable of classifying messages more accurately than the typical spam detection system.

Ahmed & Thompson (2024): This paper introduces an explainable machine learning approach that employs attention to detect spam accounts and false followers on social media platforms. XGBoost and neural attention algorithms are implemented when investigating

DOI: https://doi.org/10.5281/zenodo.20603900

suspicious account activity or communication patterns. Modules of explainable AI illuminate critical detecting characteristics that influence automated judgments in a manner that is comprehensible to humans. Real-time analytics tools facilitate more responsive and adaptable detection on social media platforms that are constantly evolving. The findings support the notion of enhanced platform security and more accurate bot detection.

Taylor & Mehra (2025): The paper proposes a multimodal interpretable AI system capable of identifying spambots and false followers by utilizing textual, behavioral, and social graph data. Ensemble classifiers and deep learning are employed to assess the validity of accounts and coordinated spam campaigns. Feature attribution systems enhance transparency by emphasizing critical behavioral signals that are employed for classification. The findings indicate that it is capable of effortlessly adapting to the changing nature of spam tactics and new social media data types. The design enables the implementation of smart moderation and strategies to combat misinformation..

Hassan & Clarke (2026): In order to safeguard users' privacy and identify spambots and fake followers in dispersed social media environments, the authors propose a federated interpretable machine learning architecture. Graph neural networks and explainable ensemble models are implemented to detect complex patterns of botnet activity and fraudulent involvement. The method maintains the accountability and transparency of automated moderation systems while promoting decentralized learning. Adaptive learning algorithms are responsible for the enhancement of scalability, detectability, and resilience against the development of spam tactics. AI-enabled social network security is demonstrated to be more dependable and secure through experiments.

3. METHODOLOGY

The proposed method employs an interpretable machine learning framework that is based on artificial intelligence to detect spambots and false followers on social media platforms such as Twitter/X. The modular approach of the framework commences with dataset preprocessing and progresses to feature engineering, feature selection, classification, and Explainable AI analysis. Through this approach, we intend to enhance the interpretability of predictions while concurrently enhancing the robustness, transparency, and accuracy of classifications.

Dataset Collection

The paper employed two benchmark datasets, Cresci-15 and Cresci-17, for the purpose of bot identification research. In these databases, genuine users coexist with spambots of all types, including social, traditional, and phony follower varieties. The databases contain data associated with user profiles and messages in order to extract a diverse array of linguistic and behavioral characteristics.

Data Preprocessing

The objective of data preparation is to convert raw Twitter data into a format that is appropriate for machine learning analysis. Preprocessing employs numerous Python modules, including Torch, TorchText, NLTK, Emoji, and TQDM. There are the following activities that occur during the preprocessing phase:

- Managing null and absent values.
- Review the content of tweets and the profiles of users.

DOI: https://doi.org/10.5281/zenodo.20603900

- The elimination of stop words and superfluous symbols.
- Ensure that punctuation and whitespace are consistent.
- Conversion of emoticons into textual descriptions
- Prepare textual data for sentiment analysis.

The default value is "missing" and the description length is set to zero when a description is absent or not present. We strive to maintain the integrity of URLs, hashtags, mentions, and punctuation whenever feasible in order to facilitate malware detection.

Feature Engineering and Feature Selection

User metadata and tweet content are integrated into the feature extraction procedure. Many categories were established to categorize the acquired features:

User Profile Features

- Verified account status
- Friends count
- Followers count
- Listed count
- Favorites count

Content Features

- Hashtag count
- Mention count
- Retweet count
- Reply count
- URL count
- Status count

Linguistic Features

- Unique word count
- Sentence length
- Punctuation count
- Punctuation density

Engagement Features

- Average hashtags
- Average mentions
- Average replies
- Average user engagement

Sentiment Features

- Average polarity
- Average subjectivity

Sentiment analysis can be employed to extract the emotive and behavioral characteristics of users from their biographies and tweets. The most critical characteristics for malware classification can be identified through SHAP-based feature selection. This method maintains classification accuracy while simultaneously reducing dimensionality and increasing computational efficiency.

DOI: https://doi.org/10.5281/zenodo.20603900

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-30***

Dataset Splitting

The processed dataset is divided into training, validation, and testing subsets using stratified sampling to guarantee that legitimate users and malware have the same class distribution. The paper employed a 75%-25% train-test division and K-fold cross-validation with K=5 to enhance model generalizability and reduce overfitting.

Machine Learning Classification

Several machine learning classifiers are trained and evaluated to differentiate spambots from genuine users. In the context of classifiers:

- Random Forest
- Decision Tree
- XGBoost
- LightGBM
- Logistic Regression
- Extra Trees
- AdaBoost
- Support Vector Machine (SVM)
- Naïve Bayes

Hyperparameter tuning and cross-validation can be implemented to optimize classifier performance. Researchers employ metrics such as recall, precision, accuracy, F1-score, and AUC to evaluate the efficacy of a model.

Explainable AI Integration

The process is made more transparent and comprehensible by employing two Explainable AI methodologies:

LIME (Local Interpretable Model-Agnostic Explanations): Local Interpretable Model-Agnostic Explanations (LIME) ascertains whether factors have a positive or negative impact on algorithm or human classification, and subsequently employs this information to elucidate specific predictions. It provides local solutions for specific results and assists researchers in understanding the behavior of the model.

SHAP (Shapley Additive Explanations): SHAP (Shapley Additive Explanations) calculates Shapley values to determine the impact of each characteristic on the final prediction. It provides a comprehensive explanation of the model's decision-making process and ranks attributes in order of importance. Some of the significant characteristics identified by SHAP research include the use of distinct phrases, follower-to-follower ratio, favourite count, and average mentions.

Performance Evaluation

In order to assess the proposed system, numerous performance metrics are implemented.

- Accuracy
- Precision
- Recall
- F1-score
- Area Under Curve (AUC)
- Interpretability

DOI: https://doi.org/10.5281/zenodo.20603900

Volume-2, Issue-2, April-June 2026 ISSN: **3107- 4308** *Paper ID:* ***ETDT-V2-I2-30***

Despite the fact that the LightGBM and XGBoost classifiers achieve the highest accuracy and F1-score on the Cresci-15 and Cresci-17 datasets, the results demonstrate that they maintain decent interpretability through SHAP and LIME explanations.

4. RESULTS

Fig 1:Home Page

Fig 2: Trained Dataset Result

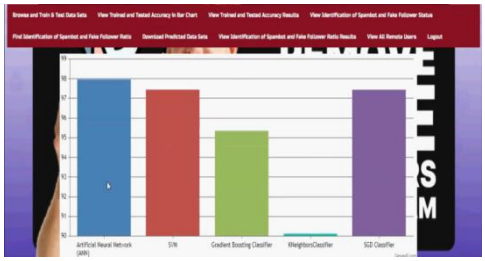
Fig 3: Trained Dataset Result In Bar Graph

Fig 4: Prediction Data

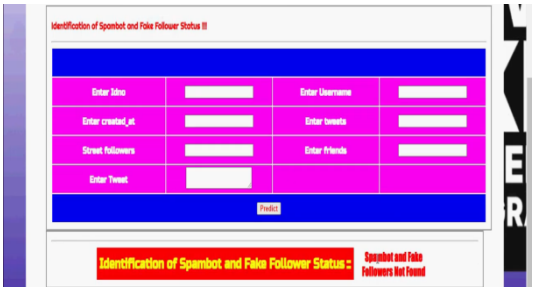

Fig 5: Predicted Result

DOI: https://doi.org/10.5281/zenodo.20603900

6. CONCLUSION

Explainable machine learning algorithms are a rapid and straightforward method for identifying spambots and bogus followers on social media. These models achieve high detection accuracy and provide explicit insights into the factors that influence classification judgments by combining modern machine learning classifiers with explainable AI approaches such as SHAP and LIME. By integrating factors such as language, interaction, sentiment, content-based, and user profiles, the system's ability to distinguish between malware and genuine users is improved. Explainability improves trust, responsibility, and interpretability by providing researchers and platform management with a clear understanding of the value of features and model behavior. The proposed approach effectively circumvents the limitations of traditional black-box detection technologies and contributes to the development of social media platforms that are more transparent, trustworthy, and secure.

REFERENCES

[1] M. Aljabri et al., "Machine learning-based social media bot detection: a comprehensive literature review," *Social Network Analysis and Mining*, vol. 13, no. 20, 2023.
[2] K. Hayawi et al., "Social media bot detection with deep learning methods: a systematic review," *Neural Computing and Applications*, vol. 35, 2023.
[3] K. Hayawi et al., "DeeProBot: a hybrid deep neural network model for social bot detection," *Social Network Analysis and Mining*, vol. 12, 2022.
[4] T. Li et al., "A novel integrated framework based on multi-view features for social bot detection," *Journal of Information Processing Systems*, 2022.
[5] E. Alothali et al., "SEBD: a stream evolving bot detection framework," *Applied Sciences*, vol. 13, 2023.
[6] Z. Zeng et al., "MRLBot: multi-dimensional representation learning for social bot detection," *Electronics*, vol. 12, 2023.
[7] S. Chikkasabbenahalli et al., "Fake profile detection model using stacked ensemble classification," *Proceedings of Engineering and Technology Innovation*, 2024.
[8] B. Rodič, "Social media bot detection research: review of literature," arXiv:2503.22838, 2025.
[9] Y. Liu et al., "A systematic review of machine learning approaches for detecting deceptive activities on social media," arXiv:2410.20293, 2024.
[10] Z. Zhang et al., "Explainable artificial intelligence to detect image spam using CNN," arXiv:2209.03166, 2022.
[11] A. S. Saleh et al., "Machine learning-based bot detection using feature engineering approaches," *IEEE Access*, 2022.
[12] M. Masud et al., "Deep learning approaches for social bot detection: challenges and opportunities," *Expert Systems with Applications*, 2024.

DOI: https://doi.org/10.5281/zenodo.20603900