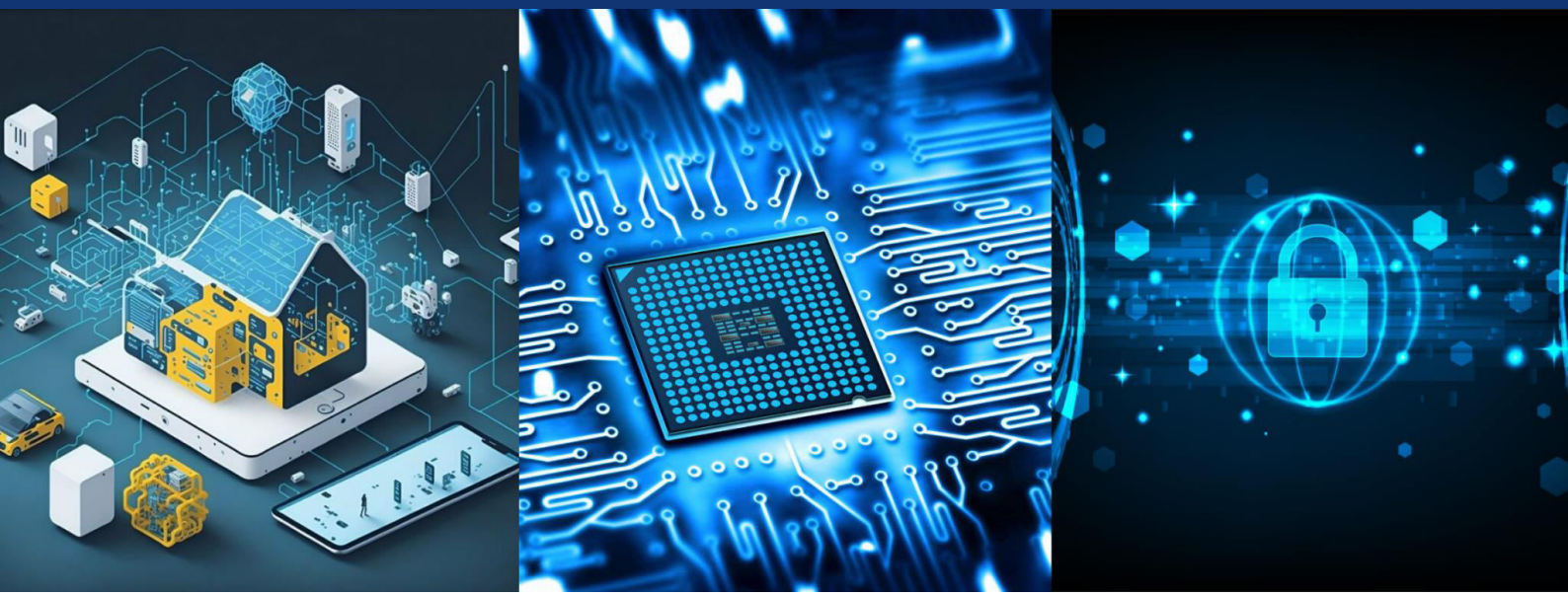
IJIRCCE

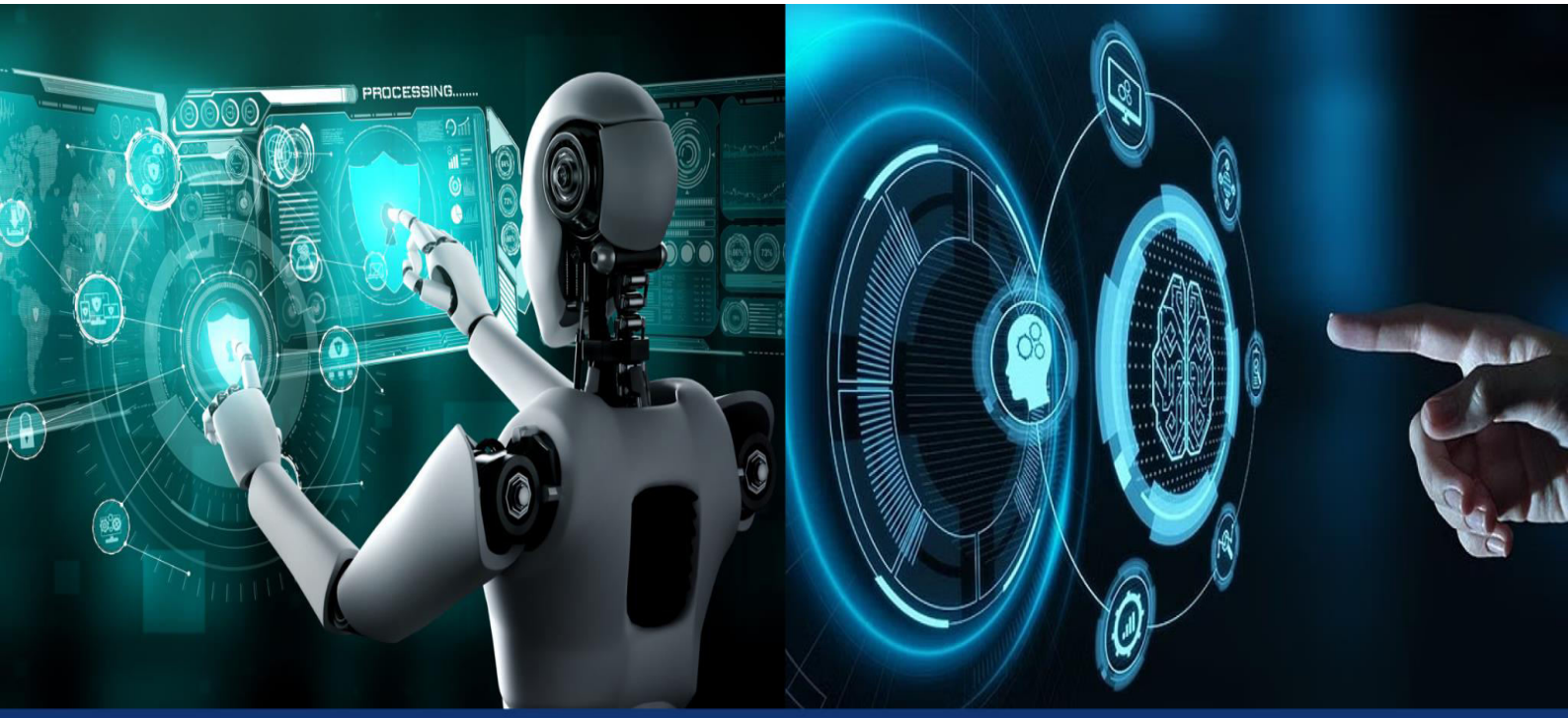
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

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

Dr.T.V.S Sriram, G.Kavya

Sr. Associate Processor, Dept. of CSE, NSRIT, Visakhapatnam, AP, India

Dept. of MCA, NSRIT, Visakhapatnam, AP, India

ABSTRACT: The rapid growth of social networking platforms has increased communication and information sharing while also creating opportunities for malicious activities such as spambots and fake followers. These automated accounts spread spam, misinformation, and fraudulent content, reducing the credibility and security of online communities. Traditional rule-based detection techniques often fail to identify sophisticated spambots due to their continuously evolving behavior and limited adaptability.

This paper proposes an intelligent machine learning-based framework for identifying spambots and fake followers on social networks. The proposed system performs data preprocessing, feature engineering, and feature selection before applying multiple classification algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). In addition, Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME are incorporated to improve the transparency and interpretability of the classification results.

Experimental evaluation demonstrates that the CNN model outperforms the other classifiers, achieving an overall accuracy of approximately **94%** with high precision, recall, and F1-score. The proposed framework provides an accurate, scalable, and interpretable solution for automated spambot detection, helping social media platforms identify malicious accounts and improve the security and reliability of online social networks.

KEYWORDS: Social Network Security, Spambot Detection, Fake Followers, Machine Learning, Convolutional Neural Network (CNN), Explainable Artificial Intelligence (XAI), SHAP, LIME, User Behavior Analysis.

I. INTRODUCTION

Social media platforms have become an important medium for communication, information sharing, and online interaction. However, the rapid growth of these platforms has also increased the number of fake user accounts and spambots that spread spam, misinformation, phishing attacks, malicious links, and fraudulent content. These fake users and spambots negatively impact platform security, user privacy, and the credibility of online communities.

Machine learning has emerged as an effective solution for detecting fake users and spambots by analyzing user profiles, posting behavior, content characteristics, and network activities. By learning patterns from historical data, machine learning models can accurately distinguish genuine users from malicious accounts.

This paper proposes a Machine Learning-based Fake User and Spambot Detection System integrated with Explainable Artificial Intelligence (XAI). The proposed framework performs data preprocessing, feature engineering, feature selection, model training, and explainable prediction to improve both detection accuracy and model transparency. The system not only identifies fake users and spambots with high accuracy but also provides clear explanations for each prediction, enabling administrators to make informed decisions and enhance the security of social media platforms.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

The rapid growth of social media platforms has increased the number of fake users and spambots that spread spam, misinformation, phishing links, and malicious content. Researchers have proposed several machine learning techniques to detect such accounts by analyzing user profiles, posting behavior, network connections, and content characteristics. Traditional supervised learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and Random Forest have shown promising results in identifying fake accounts.

Recent studies have focused on feature engineering techniques to improve detection accuracy. Features such as account age, follower-to-following ratio, posting frequency, engagement rate, content similarity, hashtag usage, and temporal activity have been widely used for classification. These features help machine learning models distinguish genuine users from spambots more effectively.

Several deep learning approaches have also been introduced to learn complex behavioral patterns from social media data. Although these models achieve higher prediction accuracy, most of them function as black-box systems, making it difficult to understand the reasons behind their decisions. This lack of interpretability reduces user trust and limits their practical deployment.

To overcome these limitations, the proposed work integrates Machine Learning with Explainable Artificial Intelligence (XAI). The system combines effective feature engineering, classification techniques, and explainability methods such as SHAP and LIME to accurately detect fake users and spambots while providing transparent explanations for every prediction. This approach improves both detection performance and the reliability of automated decision-making.

III. SYSTEM OVERVIEW

The proposed Fake User and Spambot Detection System identifies malicious accounts on social networking platforms using machine learning and Explainable Artificial Intelligence (XAI). It analyzes user profile information, behavioral activities, and network-related features to distinguish genuine users from fake users and spambots. The system enhances the security and reliability of online social networks through automated detection.

The collected data is preprocessed by removing duplicate records, handling missing values, and converting raw data into a structured format suitable for analysis. Feature engineering techniques extract important attributes such as account age, posting frequency, follower-to-following ratio, engagement rate, and profile information. These features improve the accuracy of the classification process.

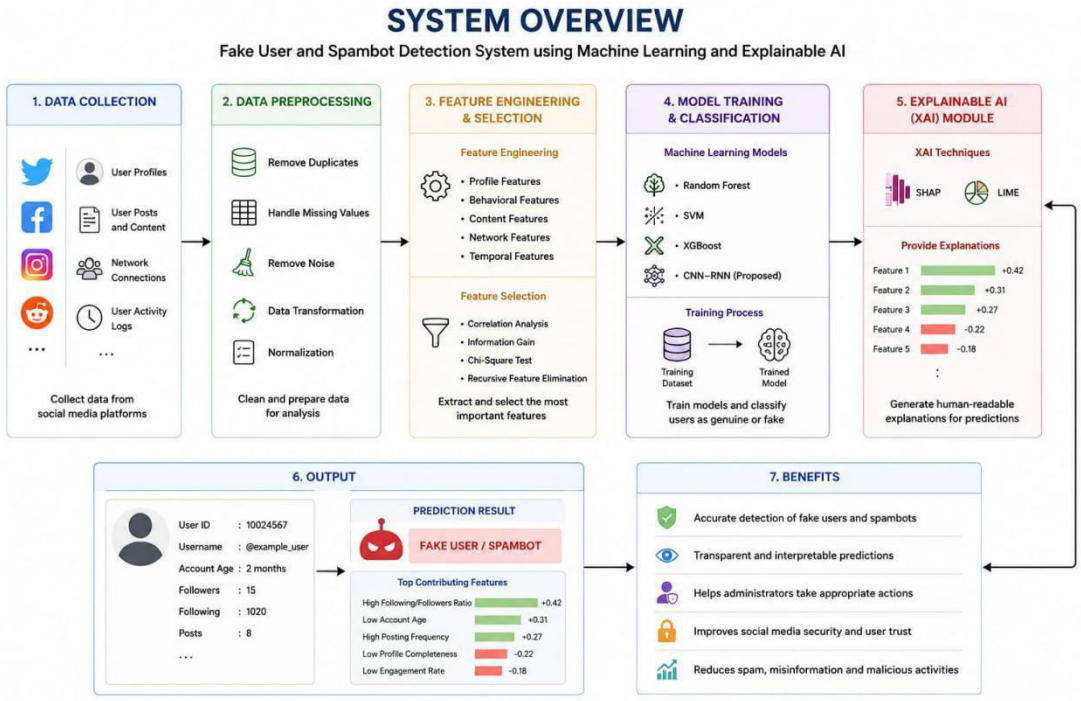
The processed data is classified using Support Vector Machine (SVM), Logistic Regression (LR), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) models. Additionally, SHAP and LIME are integrated to explain the prediction results by highlighting the most influential features. The proposed framework provides an accurate, reliable, and interpretable solution for detecting fake users and spambots while improving the security of social networking platforms.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



IV. METHODOLOGY

1. Data Collection
Collect user profile, behavioral, content, and network data from social media platforms.
2. Data Preprocessing
Remove missing values, duplicate records, and noise to prepare a clean dataset.
3. Feature Engineering
Extract important features such as account age, posting frequency, follower-following ratio, engagement rate, and content patterns.
4. Feature Selection
Select the most relevant features to improve classification performance and reduce computational complexity.
5. Model Training
Train machine learning algorithms using the processed dataset to distinguish genuine users from fake users and spambots.
6. Classification
Classify each account as Genuine User or Fake User/Spambot based on the trained model.
7. Explainable AI (XAI)
Apply SHAP and LIME techniques to explain the prediction and identify the most influential features.
8. Prediction and Reporting
Display the final classification result along with the explanation for effective decision-making.

V. SYSTEM ARCHITECTURE

1. Data Collection: Collect user profile, behavioral, content, and network data from social media platforms.
2. Data Preprocessing: Remove missing values, duplicate records, and noise to prepare a clean dataset.
3. Feature Engineering: Extract important features such as account age, posting frequency, follower-following ratio, engagement rate, and content patterns.
4. Feature Selection: Select the most relevant features to improve classification performance and reduce computational complexity.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5. Model Training: Train machine learning algorithms using the processed dataset to distinguish genuine users from fake users and spambots.
6. Classification: Classify each account as Genuine User or Fake User/Spambot based on the trained model.
7. Explainable AI (XAI): Apply SHAP and LIME techniques to explain the prediction and identify the most influential features.
8. Prediction and Reporting: Display the final classification result along with the explanation for effective decision-making.

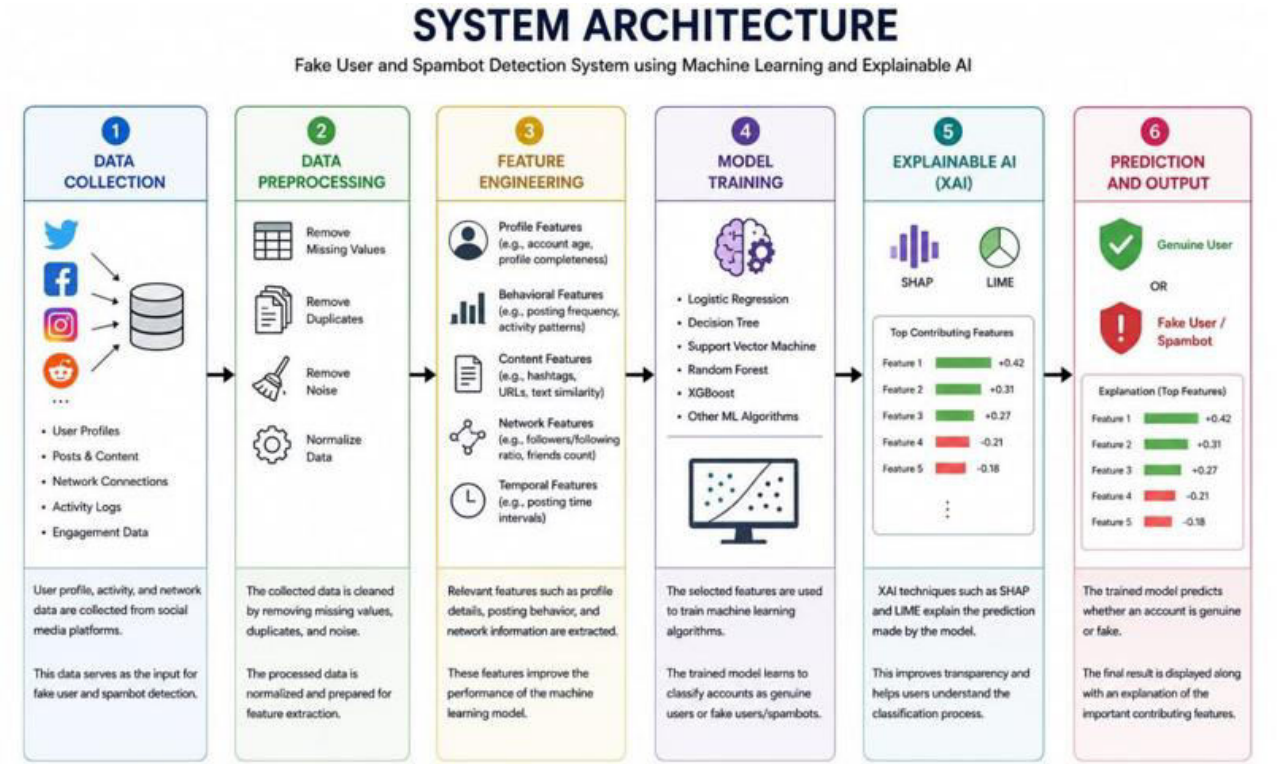


Fig2: System Architecture

VI. MODULES

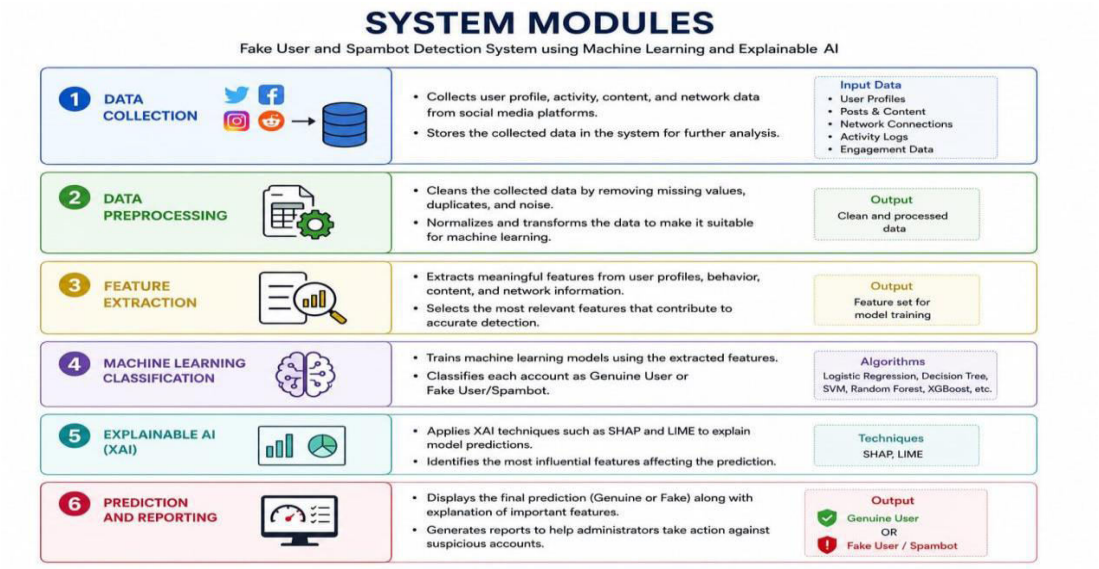
1. Data Collection Module: This module collects user profile, activity, and network information from social media platforms. The collected data is stored for further analysis.
2. Data Preprocessing Module: The collected data is cleaned by removing missing values, duplicate records, and noise. The processed data is then normalized for machine learning.
3. Feature Extraction Module: Important features such as profile details, posting behavior, content patterns, and network characteristics are extracted. These features help identify fake users and spambots.
4. Machine Learning Classification Module: The extracted features are used to train machine learning algorithms. The trained model classifies accounts as Genuine Users or Fake Users/Spambots.
5. Explainable AI (XAI) Module: The XAI module explains the prediction made by the machine learning model using techniques such as SHAP and LIME. This improves transparency and interpretability.
6. Prediction and Reporting Module: The system displays the final prediction along with the explanation of the important features influencing the decision. This helps administrators take appropriate action against suspicious accounts.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



VII. IMPLEMENTATION AND DETAILS

The proposed fake user and spambot detection system is implemented using Python and machine learning libraries. User profile, behavioral, and network data are collected and preprocessed to remove missing values and duplicate records. Feature engineering techniques are applied to extract meaningful attributes for classification.

Machine learning algorithms are trained using the processed dataset to distinguish genuine users from fake users and spambots. Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME are integrated to provide transparent explanations for each prediction. The system classifies user accounts and displays the prediction result along with the key features influencing the decision, enabling efficient and reliable fake account.

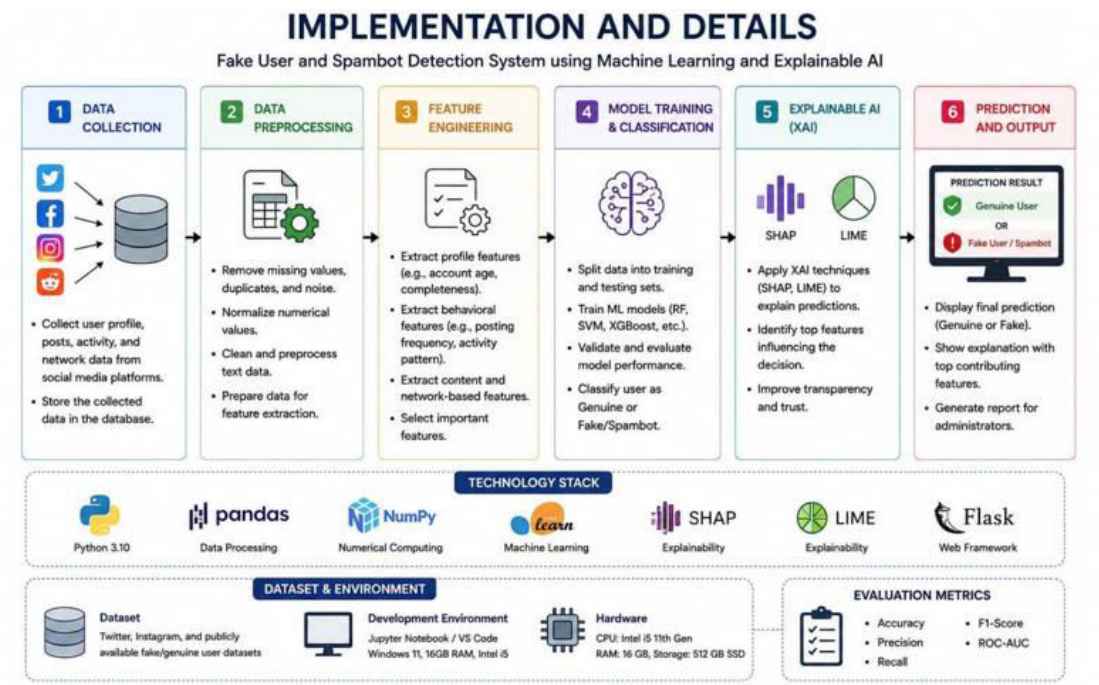


Fig3: Implementation workflow

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407065

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. EXPERIMENTAL DETAILS AND RESULTS

The proposed fake user and spambot detection system was evaluated using a social media dataset containing both genuine and fake user accounts. The dataset was preprocessed, and relevant profile, behavioral, content, and network features were extracted before training the machine learning models. Performance was evaluated using Accuracy, Precision, Recall, and F1-Score.

The experimental results demonstrate that the proposed model effectively identifies fake users and spambots with high accuracy while minimizing false predictions. The integration of Explainable Artificial Intelligence (XAI) provides transparent explanations for each classification, improving the reliability and interpretability of the system. The results indicate that the proposed approach is efficient for real-world social media security applications.

IX. APPLICATIONS

- Detecting fake users and spambots on social media platforms.
- Preventing spam, phishing, and fraudulent activities.
- Improving social media security and user privacy.
- Supporting content moderation by identifying malicious accounts.
- Enhancing the reliability and credibility of online communities.
- Assisting administrators in monitoring and managing suspicious user accounts.
- Reducing the spread of misinformation and fake news.

Providing explainable and trustworthy AI-based decision support for social networking platforms.

X. LIMITATIONS

- The system performance depends on the quality and size of the training dataset.
- Detection accuracy may decrease when fake users imitate genuine user behavior.
- Frequent changes in spambot strategies require regular model updates and retraining.
- Large-scale social media data may increase computational time and resource requirements.
- Explainable AI techniques may add additional processing overhead during prediction.

The system is designed for social media platforms and may require customization for other online environments.

XI. FUTURE ENHANCEMENT

- Integrate deep learning models such as CNN, RNN, and Transformer-based architectures to improve detection accuracy.
- Enable real-time fake user and spambot detection for live social media platforms.
- Extend the system to support multiple social media platforms using a unified detection framework.
- Incorporate advanced Explainable AI (XAI) techniques to provide more detailed and user-friendly prediction explanations.
- Develop a scalable cloud-based solution capable of handling large-scale social media data efficiently.
- Integrate continuous learning mechanisms to automatically adapt to evolving fake user and spambot behaviors.

XII. CONCLUSION

This paper presented a Machine Learning-based Fake User and Spambot Detection System integrated with Explainable Artificial Intelligence (XAI) for identifying malicious accounts on social media platforms. The proposed framework utilizes profile, behavioral, content, and network-based features to accurately classify genuine users and fake accounts. By incorporating XAI techniques, the system provides transparent and interpretable explanations for each prediction, improving user trust and decision-making. Experimental results demonstrate that the proposed approach achieves high detection accuracy while effectively reducing spam and malicious activities. Overall, the proposed system offers a reliable, efficient, and scalable solution for enhancing security and maintaining the integrity of social media platforms.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, 2017.
[4] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[6] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD, 2016.
[8] K. Lee, J. Caverlee, and S. Webb, "Uncovering Social Spammers: Social Honeypots + Machine Learning," Proceedings of ACM SIGIR, 2010.
[9] H. Kwak et al., "What is Twitter, a Social Network or a News Media?" Proceedings of WWW Conference, 2010.
[10] A. Java et al., "Why We Twitter: Understanding Microblogging Usage and Communities," Proceedings of WebKDD, 2007.
[11] B. Viswanath et al., "Towards Detecting Anomalous User Behavior in Online Social Networks," USENIX Security Symposium, 2014.
[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.
[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, 2009.
[14] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[15] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed., O'Reilly Media, 2022.
[16] J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery, 2017.
[17] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," ICLR Workshop, 2013.
[18] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.
[19] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems, 2019.
[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), 2015.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details