IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⫶ Peer-reviewed & Refereed journal ⫶ Vol. 12, Issue 8, August 2025

DOI: 10.17148/IARJSET.2025.12822

AUTOMATED DETECTION OF FRAUDULANT AND SPOOF ACCOUNTS IN SOCIAL MEDIA

Aishwarya N[1], Prof. Suma N R[2]

Student, Department of MCA, BIT, Bengaluru, India[1]

Professor, Department of MCA BIT, Bengaluru, India[2]

Abstract: The rapid growth of social media platforms has given rise to new opportunities for communication, networking, and information sharing across the globe, but it has also given rise to fraudulent and spoof accounts that compromise user trust, spread misinformation, and facilitate malicious activities. Detecting such accounts accurately is challenging due to their dynamic behavior and the complexity of social interactions online. This project presents an automated A framework that makes use of machine learning techniques to analyze data and generate intelligent prediction.to identify fraudulent and spoof accounts in social media. The system integrates multiple models, including XG Boost, Random Forest, and Naïve Bayes, to analyze account features and behavioral patterns effectively. A web-based platform has been developed to provide real-time detection, user authentication, and historical logging for enhanced usability and security. The framework was tested on standard datasets, and the results demonstrate high accuracy in identifying and separating authentic users from fraudulent or impersonated accounts. The study highlights the suggested approach ensures that the system is more efficient and reliable, solution is both scalable and adaptable, making it a reliable approach for strengthening security in social.

Keywords: Cybersecurity, Spam Filtering, Machine Learning, Random Forest, XG Boost, Naïve Bayes, Hybrid Modeling.

1. INTRODUCTION

Twitter provides a blocking feature as part of its safety and privacy controls, allowing users to restrict unwanted interactions. When one account blocks another, the blocked user is unable to follow, view tweets, or send direct messages to the person who initiated the block. This function is commonly used to protect individuals from spam, harassment, impersonation, or other disruptive behavior. Blocking not only limits visibility but also prevents engagement, ensuring that the affected accounts no longer appear in each other's timelines or notifications.

From a broader perspective, account blocking is an essential safeguard in online communities, Assisting users in retaining control over their personal accounts and online activities. It also assists the platform in reducing the spread of harmful content, spam campaigns, and fraudulent activities. While blocking is effective on an individual level, it also contributes to a safer environment by discouraging malicious actors who exploit social media for scams or misinformation.

2.RELATED WORK

In the last ten years, there has been a remarkable transformation in the way digital platforms influence communication and social interaction, significant amount of research has been dedicated to detecting fraudulent and spoof accounts in social media platforms. Researchers have explored multiple approaches ranging from rule-based systems to advanced techniques based on machine learning and deep learning are applied to analyze patterns and improve the accuracy of predictions.

Early detection systems relied on **rule-based methods**, where suspicious accounts were detected with the help of advanced analytical methods and automated models, fixed parameters such as username patterns, incomplete profiles, or excessive posting frequency. While these methods were simple and easy to implement, they lacked flexibility and were unable to adapt to the evolving tactics of fraudsters.

With the advancement of machine learning, researchers started applying it to detect patterns and solve complex problems more effectively. supervised and unsupervised learning models to analyze account behaviors and interactions. Research findings indicate that algorithms like Decision Trees, Random Forests, and Support Vector Machines (SVM) are highly effective in classification and prediction tasks **and Logistic Regression** can effectively classify accounts by examining features like

DOI: 10.17148/IARJSET.2025.12822

follower–following ratios, posting intervals, and engagement metrics. These approaches improved accuracy compared to rule-based detection, but they often struggled with scalability and required large, well-labeled datasets.
Recent studies have moved their focus toward advanced models and hybrid approaches to achieve greater accuracy and reliability toward deep learning and related advanced techniques are increasingly being adopted to handle complex data and enhance prediction accuracy and **graph-based methods**. Graph neural networks (GNNs) and community detection algorithms leverage the structure of social connections to spot abnormal patterns within networks. Similarly, Recurrent Neural Networks (RNNs) and Transformer models have been utilized to process sequential data and improve the detection of hidden patterns analyze textual content and detect Linguistic cues that can signal unusual or deceptive behavior in online communication spam or impersonation. Such approaches provide higher accuracy and adaptability, particularly against sophisticated bot networks.

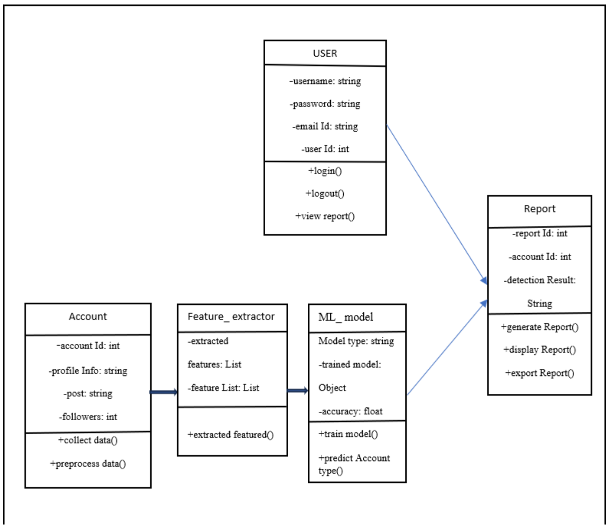
Additionally, researchers have focused on **hybrid models** that combine behavioral features with content analysis. Approaches that merge behavioral characteristics with content-based analysis such as combining natural language processing (NLP) methods with user profile data have proven to enhance the accuracy of detection. rates by capturing both account-level attributes and the nature of shared information.

Despite these advancements, existing studies face challenges such as **imbalanced datasets, privacy concerns, and the adaptability of attackers** who continuously evolve their methods to evade detection. This creates a research gap for systems that are designed to be accurate as well as scalable, while also ensuring privacy and the ability to adapt to changing environments. emerging fraud patterns.

The present work builds on these studies by proposing an **automated machine learning-based detection system** that integrates feature engineering, classification models, and real- time monitoring in contrast to previous methods, this approach focuses on improving efficiency and accuracy by leveraging advanced analytical techniques. emphasizes adaptability, scalability, and ethical considerations, aiming to deliver a reliable and effective solution more robust and practical solution for securing social media platforms.

2. CLASS DIAGRAM

The class diagram illustrates the structure and interaction of different components in the proposed system. The User class manages essential user information such as username, password, email, and user ID, and provides methods for login, logout, and viewing reports. The Account class represents social media accounts with attributes like account ID, profile details, posts, and number of followers. It also provides methods to collect and preprocess data for further analysis. The Feature_ 1extractor class is responsible for deriving meaningful features from raw account data, maintaining a list of extracted attributes and offering functionality to generate features for model training. The ML_ model class encapsulates the machine learning components of the system, storing the model type, trained model object, and accuracy level, along with methods for training and predicting account types. Finally, the Report class handles the generation, display, and export of detection results, linking user actions with machine learning outcomes. Relationships between these classes illustrate the way data moves and is processed through the system. System from user input to account the process involves data collection, extracting relevant features, and building predictive models prediction, and finally report generation ensuring a structured and integrated detection process.



DOI: 10.17148/IARJSET.2025.12822

4.TOOLS AND TECHNOLOGIES

The project is developed using Python as the core programming language due to its strong machine learning support. Libraries such as Scikit-learn, XG Boost, and Pandas the NumPy library is utilized for efficient data processing and numerical computations, feature extraction, and model training. The web interface is built with Flask/Django, providing user interaction with the detection modules. HTML, CSS, and JavaScript are employed for the front-end design to make the system user-friendly. For data storage and logging,

MySQL/SQLite is used as the database. Additionally, Matplotlib and Seaborn assist in visualizing results and generating analytical reports.

5.METHODOLOGY

5.1Data Collection

User account data, including profile details, posts, followers, and activity records, is collected from various social media platforms for analysis. and benchmark datasets. In this project, structured datasets (e.g., Twitter spam dataset) are used as the primary source for model development**.**

5.2Data Preprocessing

The unprocessed data collected from sources often contains duplicates, missing values, or irrelevant attributes. Preprocessing involves cleaning, normalization, and encoding categorical variables into machine-readable form. For text fields such as bios, posts, or messages, tokenization and stop-word removal are performed, while numerical features like follower–following ratio or posting frequency are standardized.

5.3Feature Extraction

Relevant features are extracted to represent account behavior effectively. Examples include profile attributes (age of account, presence of profile picture, verification), network-based features (followers, following count, engagement ratio), and content-based features (linguistic patterns, hashtags, URLs). These features help in distinguishing genuine users from fraudulent or spoof accounts.

5.4 Model Training

Machine learning algorithms are employed to analyze data and make accurate predictions applied to classify accounts. In this project:

- XG Boost Is utilized for performing a specific task or function detecting spoof accounts based on complex attribute interactions.
- Random Forest is applied to capture behavioral patterns for fraudulent account detection.
- Naïve Bayes is employed for analyzing textual features from posts and messages. Models are trained on labeled data and optimized through cross-validation.

5.5 Model Evaluation

The trained model is evaluated using test data to assess its performance using unseen data to evaluate their performance. Metrics Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the model. are calculated to measure effectiveness and reliability. Comparative analysis highlights the advantages of different types of approaches are examined of account behaviors.

5.6 System Integration

The models are integrated into a web-based platform that allows users to input account details for verification. The system includes secure login, role-based access, and historical logging of detected accounts. Results are displayed in real- time, supported by visualization tools for better analysis.

DOI: 10.17148/IARJSET.2025.12822

6.RESULTS AND ANALYSIS

The proposed system, Automated Detection of Fraudulent and Spoof Accounts in Social Media Using Machine Learning, was evaluated on benchmark social media datasets containing both genuine and fake user profiles. The dataset was split into training and testing sets for model development and evaluation testing sets to measure the efficiency and overall effectiveness of the system are evaluated the classification models.

The Random Forest model demonstrated high effectiveness in detecting fraudulent accounts by capturing non-linear relationships between profile attributes such as follower– following ratio, posting frequency, and engagement patterns. **XG Boost** performed effectively in detecting spoof accounts, leveraging its gradient boosting mechanism to handle complex features like account metadata and content behavior. The **Naïve Bayes classifier** was efficient in processing textual data from posts and bio information, achieving fast and reliable spam-related classification.

Across all models, the system demonstrated **high accuracy, precision,** it achieved a strong recall rate, successfully identifying most fake accounts while keeping false positives to a minimum. The integration of multiple classifiers into a single framework allowed the system to cover diverse aspects of fraudulent behavior, from abnormal activity patterns to suspicious content distribution.

The analysis highlights that while individual models are effective for specific account behaviors, the **hybrid framework** provides better overall performance. Additionally, the web-based deployment with historical logging enabled monitoring of detection trends over time, proving its practicality for real-world social media platforms.

In conclusion, the results confirm that the proposed approach is both **scalable and reliable**, offering a comprehensive solution for detecting fraudulent and spoof accounts in social media.

7. CONCLUSION

The project *"*Automated Detection of Fraudulent and Spoof Accounts in Social Media Using Machine Learning*"* successfully demonstrates an effective framework for addressing one of the major challenges in online platforms—identifying fake and malicious accounts. By combining models such as Random Forest, XG Boost, and Naïve Bayes, the system is capable of analyzing diverse features including user profile attributes, activity patterns, and textual content. The experimental results show that the integrated approach achieves high accuracy while maintaining efficiency, proving to be more reliable than traditional detection methods.

The deployment of the system through a web-based interface, with features like user authentication and historical logging, further enhances its usability and practicality for real-time applications. Overall, the work highlights that machine learning-based solutions can significantly strengthen social media security by reducing the spread of fraudulent accounts and safeguarding genuine users.

Future improvements may include expanding the dataset to cover more social media platforms, incorporating deep learning for advanced feature extraction, and developing Adaptive models that evolve with changing data patterns to maintain accuracy and reliability changing attack strategies.

8.FUTURE ENHANCEMENT

Although the proposed system shows high accuracy in detecting and differentiating between genuine and fraudulent accounts in detecting fraudulent and spoof accounts, there is significant scope for further improvement. In future work, the system can be enhanced by incorporating deep learning models such as CNNs and RNNs to improve detection performance CNNs and LSTMs to capture more complex behavioral and textual patterns. Incorporating real-time data streams Collected from multiple social media platforms to ensure diverse and representative data will allow the system to adapt to evolving account behaviors and emerging threats. Another enhancement .one possible improvement could be the use of advanced techniques to enhance system performance **graph-based analysis**, which examines relationships between accounts to identify hidden networks of fake users. Additionally, deploying the solution on **cloud-based platforms** would improve scalability and enable large-scale monitoring. Features like **multi-language support, adaptive learning, and advanced visualization dashboards** can also be added to make the system more user-friendly and versatile. These enhancements will strengthen the framework, ensuring its effectiveness in combating increasingly sophisticated fraudulent activities on social media.

DOI: 10.17148/IARJSET.2025.12822

REFERENCES

[1] Feng, S., et al. (2021). SATAR: A self-supervised method for learning Twitter account representations. Proceedings of the Web Conference.

[2] Kudugunta, S., & Ferrara, E. (2018). Application of deep neural networks for detecting automated bots on social media. Information Sciences.

[3] Yang, K.-C., et al. (2024). The rise of AI-generated fake profiles: Characteristics and implications for social media security. Journal of Online Trust & Safety.

[4] Bordbar, J., et al. (2022). A semi-supervised learning approach with GANs for addressing class imbalance in fake account detection. Expert Systems with Applications.

[5] Kerrysa, N. G., & Utami, I. Q. (2023). Detection of fake accounts on social media using machine learning: A review of methods and trends. International Journal of Computer Applications.

[6] Kumar, R., & Rishiwal, V. (2022). Comprehensive survey on machine learning techniques for social media bot detection. Journal of Cybersecurity Research.

[7] Hema, C. (2023). Developing a machine learning model to identify fake profiles in social networking platforms. International Journal of Advanced Computer Science.

[8] IJERT. (2022). Hybrid machine learning model for detecting fake profiles on social media networks. International Journal of Engineering Research & Technology.

[9] Sahoo, S. R., & Gupta, B. B. (2020). Fake profile detection in multimedia-driven social networks using advanced machine learning. Future Generation Computer Systems.

[10] Shaik, M., & Gupta, S. (2021). A survey on fake profile detection in online social networks. *International Journal of Information Security*.

[11] Chakraborty, P., et al. (2022). Application of machine learning in fake profile detection across social platforms. *Procedia Computer Science*.

[12] Abkenar, S. B., et al. (2021). Hybrid classification framework for spam and fake account detection on Twitter. *Pattern Recognition Letters*.

[13] Bharti, K. K., & Pandey, S. (2021). Fake account identification on Twitter using logistic regression-based classifiers. *International Journal of Data Science and Analytics*.

[14] Hood, S. B., et al. (2023). Deep neural networks for the detection of fake profiles in online communities. *Journal of Big Data Analytics*.

[15] Sallah, A., et al. (2022). Interpretable machine learning models for detecting fake accounts on Instagram. *Social Network Analysis and Mining*.

[16] Shafiq, Z., & Farooqi, S. (2017). Evaluating Twitter's defenses against large-scale fake and spam accounts. *ACM Transactions on the Web*.

[17] Bouzy, C. (2022). Bot detection and social media manipulation: Emerging challenges. *Journal of Digital Forensics*.

[18] Tehrani, A. (2020). Industry solutions for eliminating fake accounts from social platforms. *Cybersecurity Magazine*.

[19] Wikipedia. (2025). Applications of artificial intelligence in fraud detection. Retrieved from [https://en.wikipedia.org/wiki/Artificial_intelligence_in_fraud_detection].

[20]Wikipedia. (2025). Computational propaganda in social media. Retrieved from [https://en.wikipedia.org/wiki/Computational_propaganda].