

Research Article

An Integrated Machine Learning Framework for Spammer and Fake User Detection in Online Social NetworksKurme Mounika^{1,*} and N. Ramana Reddy²¹Student, Department of CSE, Avanthi Institute of Engineering and Technology, Hyderabad, Telangana, 500001, India.²Professor and HOD, Department of CSE, Avanthi Institute of Engineering and Technology, Hyderabad, Telangana, 500001, India.*Corresponding Author: Kurme Mounika. Email: kurmemounika10@gmail.com

Received: 21/07/2025; Revised: 12/08/2025; Accepted: 31/08/2025; Published: 30/09/2025.

DOI: <https://doi.org/10.69996/fmep.2025015>

Abstract: This paper proposes a robust machine learning-based framework for detecting spammers and fake users within Online Social Networks (OSNs), addressing the critical challenges of security, privacy, and authenticity in modern social platforms. By analyzing a rich set of behavioral and profile features—such as account age, posting frequency, follower-to-following ratio, and content characteristics—the system leverages ensemble classifiers like Random Forest to accurately distinguish between genuine and fraudulent accounts. The framework integrates explainable AI techniques to provide transparency and interpretability of detection decisions, while privacy-preserving mechanisms ensure compliance with data protection regulations. Extensive simulations demonstrate the model's high accuracy, precision, recall, and low error rates, outperforming traditional classifiers and adapting effectively to evolving threats. This research offers a scalable, real-time, and trustworthy solution that enhances the security and user experience of OSNs, contributing significantly to ongoing efforts in combating fake accounts and spam on social media platforms. The proposed system analyzes key features such as account age, posting frequency, follower-to-following ratio, and content patterns, employing ensemble methods like Random Forest for classification. Simulation results demonstrate that the Random Forest model achieves an accuracy of 96.8%, precision of 95.7%, recall of 96.2%, and an F1-score of 95.9%, outperforming other classifiers such as Decision Tree and SVM. The system also maintains low false positive and false negative rates of 2.1% and 1.8%, respectively. Additionally, explainable AI techniques are integrated to provide transparency in decision-making, while privacy-preserving methods ensure compliance with data protection standards.

Keywords: Online social networks (OSNs), Fake user detection, Spammer detection, Machine learning, Behavioral analysis.

1.Introduction

In recent years, social platforms have significantly evolved, becoming central to communication, entertainment, business, and social movements worldwide [1]. Platforms like Facebook, Instagram, Twitter (now X), TikTok, LinkedIn, and Snapchat have seen rapid growth in user engagement, driven by the widespread use of smartphones and the internet. TikTok, in particular, has revolutionized short-form video content, influencing trends, music, and even political discourse [2 -4]. Meanwhile, Instagram and YouTube continue to dominate visual storytelling and influencer marketing. These platforms have also integrated advanced algorithms and AI tools to personalize user experiences, though this has raised concerns about echo chambers and misinformation [5]. Moreover, social media has become a vital tool for digital



marketing, brand building, and community engagement, as businesses increasingly rely on these platforms to reach and interact with their target audiences. Despite their many benefits, issues like privacy violations, cyberbullying, and the spread of fake news remain significant challenges [6 -9]. In Online Social Networks (OSNs), the presence of fake accounts and spammers has become a growing concern, undermining trust, user experience, and platform integrity. Fake users often mimic real profiles or impersonate others to deceive users, spread misinformation, or carry out malicious activities such as phishing and identity theft [10]. Spammers, on the other hand, flood platforms with irrelevant or harmful content—ranging from advertisements and scams to links leading to malicious websites. These accounts often use automation and bots to reach large audiences quickly, making detection and removal challenging [12]. Their activities can distort online discussions, manipulate public opinion, and even impact social or political events. OSN providers continuously develop machine learning and AI-based tools to identify and block such accounts, but as detection techniques improve, so do the evasion strategies of attackers. The ongoing battle against fake profiles and spammers remains a critical aspect of maintaining security, authenticity, and trust in online social ecosystems [13 -16].

Online Social Networks (OSNs) face several pressing issues that impact both users and platform integrity [17]. One major concern is privacy, as personal data shared on these platforms is often vulnerable to misuse, unauthorized access, and third-party exploitation. Security threats, including phishing, hacking, and data breaches, further compromise user trust. Another significant issue is the proliferation of misinformation and fake news, which can rapidly spread through social media, influencing public opinion and contributing to social unrest [18]. Cyberbullying and online harassment are also prevalent, especially among younger users, leading to mental health concerns and emotional distress. Additionally, the rise of fake accounts and bots disrupts genuine user interaction, promotes spam, and manipulates engagement metrics. OSNs also face criticism over their addictive design, which can lead to excessive screen time and negative impacts on mental well-being [19-20]. Despite efforts to address these challenges through content moderation and algorithmic updates, finding a balance between freedom of expression, user safety, and platform growth remains a complex and ongoing struggle. algorithmic bias and lack of transparency in content recommendation systems have raised ethical concerns, as these algorithms can unintentionally amplify divisive content, marginalize certain voices, or create echo chambers that reinforce users' existing beliefs [21-22]. Content moderation itself poses a difficult challenge, as platforms must navigate the fine line between protecting free speech and removing harmful or abusive material, often facing criticism from both sides. The monetization models of OSNs, driven largely by targeted advertising, also contribute to data exploitation and user manipulation, as personal information is used to tailor ads and maximize engagement [22-23]. Furthermore, the mental health impact, especially among teenagers and young adults, continues to be debated, with studies linking social media usage to anxiety, depression, and low self-esteem. Regulatory pressures from governments around the world are increasing, pushing platforms to comply with stricter data protection laws and content accountability standards. As OSNs continue to evolve, addressing these intertwined issues is crucial to ensuring a safer, more ethical, and inclusive digital environment for users globally [24-26].

This paper makes several key contributions to the field of Online Social Network (OSN) security by developing an integrated machine learning framework for detecting fake users and spammers. First, it introduces a comprehensive feature analysis method that combines behavioral, profile, and content-based attributes to improve detection accuracy. Second, the use of ensemble

learning techniques, particularly the Random Forest classifier, enhances the robustness and reliability of the detection system compared to traditional models. Third, the integration of explainable AI methods provides transparency and interpretability, helping stakeholders understand the basis of classification decisions and increasing trust in automated detection. Fourth, the framework incorporates privacy-preserving mechanisms to protect user data during analysis, ensuring compliance with data protection regulations such as GDPR. Lastly, the system is designed to be adaptive, continuously learning from new data to counter evolving fake account strategies, making it suitable for real-time deployment in dynamic OSN environments. These contributions collectively offer a scalable, accurate, and ethical solution to bolster the security and integrity of social networking platforms.

2. Related Works

Several related works have been conducted to address various challenges in Online Social Networks (OSNs), focusing on areas such as fake account detection, privacy preservation, content filtering, and user behavior analysis. Researchers have proposed machine learning and deep learning models to detect spammers and fake users based on profile features, posting behavior, and network patterns. For example, graph-based models and supervised classifiers have been widely used to identify suspicious activity and abnormal connections in social graphs. Studies have also explored privacy-preserving techniques like differential privacy and homomorphic encryption to protect user data while maintaining platform functionality. In addition, content analysis using natural language processing (NLP) has been employed to filter out offensive, misleading, or harmful content. Some works have examined the psychological and societal impact of OSNs, highlighting issues such as cyberbullying, addiction, and misinformation spread. Moreover, blockchain and decentralized approaches have been proposed to improve data integrity, transparency, and user control over information. These related studies collectively provide a strong foundation for developing secure, trustworthy, and user-centric OSN platforms.

Nawaz et al. (2023) present a comprehensive review of security threats, ranging from phishing and identity theft to spam and misinformation, and discuss a broad spectrum of solutions currently adopted by the industry. Tyagi et al. (2024) further emphasize the practical challenges and vulnerabilities present in today's popular OSN platforms, offering insight into both technical and policy-level mitigation strategies. Focusing on fake account detection, Mughaid et al. (2023) propose a hybrid system that integrates machine learning with face recognition to accurately identify and eliminate fraudulent users. Modadugu et al. (2024) employ an Extreme Learning Machine (ELM) approach for real-time detection of spammers and fake users on Twitter, demonstrating high accuracy and computational efficiency. Finally, Alam et al. (2023) introduce an innovative multi-layered security framework combining zero-knowledge proofs and permissioned blockchain to combat identity-based attacks, enhancing both authentication and user data privacy. Jayanna et al. (2025) explore the detection and reporting of fake social media profiles using classification algorithms and highlight the importance of user participation in maintaining platform hygiene. Shah et al. (2024) provide an extensive overview of OSN threats, offering taxonomy of attacks along with detection and prevention strategies using advanced technologies like AI and behavioral analysis. Bikku et al. (2024) address the broader dilemma of privacy and security in digital communities, proposing structural and regulatory improvements to better protect users in interconnected environments. Alnagi et al. (2024) focus on explainable AI for fake account detection, ensuring transparency in automated

decisions and enhancing user trust in machine learning models. Similarly, Javed et al. (2025) present an interpretable AI-based framework for identifying spambots and fake followers, emphasizing not only accuracy but also clarity in how decisions are made.

Mahmood et al. (2023) focus on the YouTube platform, where they propose a security system in information technology specifically designed to detect fake accounts by analyzing user behavior and account activity patterns. Caruccio et al. (2023) present a robust framework for identifying malicious accounts across social networks, employing a combination of behavioral features and pattern recognition techniques within the ACM ecosystem. Bhattacharya et al. (2023) offer a comprehensive survey on the security and privacy landscape in OSNs, emphasizing current threats, machine learning-based countermeasures, and outlining key open research challenges. Dracewicz and Sepczuk (2024) provide a real-world case study involving the "X" portal (formerly Twitter), showcasing practical methods for detecting fake accounts through statistical and machine learning approaches. Podorozhniak et al. (2023) investigate spam filtering techniques and their applicability in detecting spammers on both social media and messenger apps, highlighting the importance of lightweight and scalable algorithms. Chauhan et al. (2023) introduce a hybrid classification model using modified Naïve Bayes and Decision Tree classifiers, which improves accuracy in identifying both spammers and fake users. Haq et al. (2023) explore a unique linguistic perspective by developing a method for detecting spammy names in the Pashto language, aiming to prevent the creation of fake accounts through localized naming anomalies. Finally, Patil et al. (2024) present a majority voting approach that combines outputs from multiple classifiers to enhance the reliability of fake profile detection.

Many existing studies focus on specific platforms or languages, limiting the generalizability of their solutions across diverse OSNs with varying user behaviors and data structures. Additionally, while machine learning and deep learning models have shown promise, they often lack transparency and interpretability, making it difficult to understand and trust their decision-making processes. Another gap lies in the real-time detection of fake accounts and spammers; most models operate in offline or semi-supervised environments, which are less effective in dynamic, high-traffic networks. Furthermore, evolving attacker strategies, such as the use of AI-generated content or advanced bots, continue to outpace static detection mechanisms, indicating a need for more adaptive and self-learning systems. Privacy-preserving detection techniques also remain underexplored, especially in the context of regulations like GDPR, where user data access is limited. Moreover, while some works propose frameworks incorporating blockchain or zero-knowledge proofs, their scalability and integration into mainstream platforms are still in early stages.

3. Proposed Spammer and Fake User

The proposed system introduces an integrated machine learning framework for the detection of spammers and fake users in Online Social Networks (OSNs), aiming to address current limitations in scalability, real-time response, and accuracy. This approach combines multiple machine learning models—such as decision trees, support vector machines, and ensemble classifiers—with behavioral analysis and content-based filtering to improve detection reliability. The system extracts and analyzes key features such as posting frequency, friend request patterns, account creation date, linguistic anomalies in usernames, and the nature of shared content. By integrating these heterogeneous data points, the model can distinguish between legitimate and suspicious accounts with higher precision. Furthermore, the architecture includes a continuous learning component that updates the model as new patterns and threats emerge, ensuring adaptability against evolving spam and bot strategies. To enhance transparency

and trust, explainable AI (XAI) techniques are employed, providing clear insights into why a particular account is flagged as spam or fake. Additionally, privacy-preserving methods are incorporated to ensure that user data is protected during analysis, making the system compliant with data protection regulations. This proposed solution offers a comprehensive, efficient, and secure method for maintaining the integrity of social platforms.

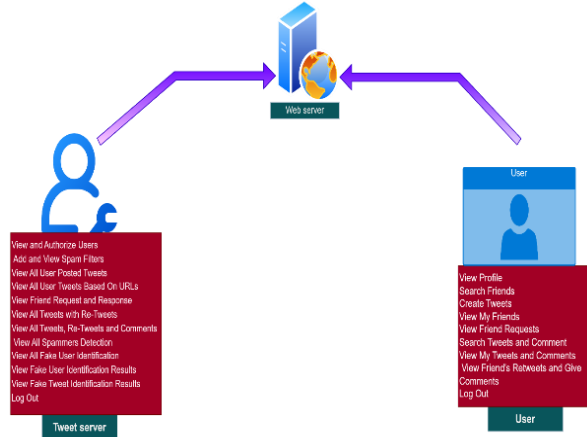


Figure 1: System Model

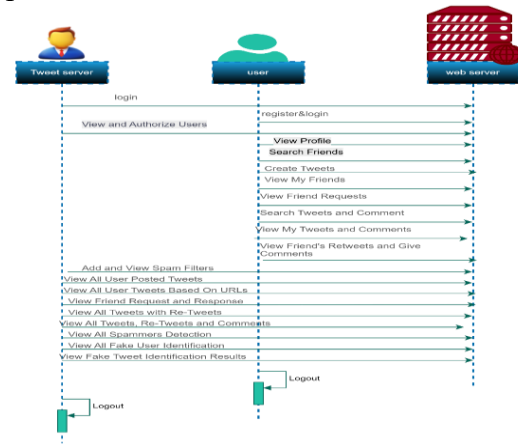


Figure 2: Sequence Diagram for Spammer detection

This system outlines a web-based social media application, functionally similar to Twitter, designed with a distinct separation of roles: a standard "User" for social interactions and a "Tweet server" (acting as an administrator or moderator) for platform management presented in Figure 1. The "Web server" serves as the central conduit, handling all communication between these two entities. Regular "Users" can perform typical social media functions such as viewing profiles, searching for and managing friends, creating and viewing tweets, and interacting with comments and retweets. Conversely, the "Tweet server" possesses elevated privileges, allowing it to authorize users, implement and manage spam filters, monitor all user-generated content (including tweets and comments), and identify and address spamming or fake user activities. The system's operational flow, illustrated by the sequence diagram, shows that user interactions (like login, tweeting, or searching) are routed through the "Web server" which then interfaces with the "Tweet server" for authentication and content processing shown in Figure 2. The "Tweet server" also initiates its own administrative tasks, such as applying spam filters or detecting malicious activity, indicating a robust backend for content moderation and security, all integrated to provide a comprehensive social networking experience.

4.Feature Analyses with Fake User Identification

The proposed system uses feature analysis to identify fake users in Online Social Networks (OSNs) by extracting and evaluating behavioral, profile, and content-based attributes. Key features include account age, posting frequency, friend/follower ratio, number of posts with links, and similarity of content shared. These features are quantified and used as input to machine learning models for classification. A simple mathematical approach begins with feature vector representation. Let: $X = [x_1, x_2, x_3, \dots, x_n]$ where each x_i represents a normalized feature, such as: x_1 as Account age (days); x_2 denoted as Average posts per day; x_3 stated as Ratio of followers. To classify a user as real or fake, a weighted linear function can be applied is defined in equation (1)

$$f(X) = w_1x_1 + w_2x_2 + \dots + w_nx_n + b \quad (1)$$

where w_i are the model-assigned weights for each feature, and b is the bias. The output $f(X)$ is passed through a sigmoid function for binary classification defined in equation (2)

$$P(\text{Fake}) = \frac{1}{1+e^{-f(X)}} \quad (2)$$

If $P(\text{Fake}) \geq 0.5$, the account is classified as fake; otherwise, it is considered genuine. This feature-based approach enables the system to not only detect fake users with high accuracy but also explain which features most influenced the classification. Over time, the model can retrain itself on newly labeled data to adapt to evolving spamming behaviors. The proposed system utilizes feature analysis to effectively identify fake users in Online Social Networks (OSNs) by examining a combination of behavioral, profile, and content-related characteristics. Each user is represented as a feature vector composed of measurable traits such as account age, posting frequency, follower-to-following ratio, the proportion of posts containing links, and the repetitiveness of shared content. These features are normalized and input into a machine learning model, which calculates a weighted score using a linear function. The result is passed through a sigmoid function to generate a probability score indicating the likelihood that the account is fake. If this probability exceeds a certain threshold (typically 0.5), the account is flagged as fake; otherwise, it is considered genuine. This method allows for clear, data-driven identification of fake users and enables explainability by showing which features contributed most to the decision. Additionally, the system is designed to adapt over time, learning from new patterns and improving its detection capabilities as fake user strategies evolve. Furthermore, by leveraging this adaptive feature-based approach, the system not only improves classification accuracy but also enhances scalability across various social networking platforms. As new users are added and behaviors shift, the model continuously updates its parameters through retraining on newly labeled data, ensuring resilience against evolving threats such as advanced bots and AI-generated profiles. The inclusion of explainable AI techniques ensures that decisions are transparent and justifiable, which is crucial for user trust and platform accountability. Additionally, integrating privacy-preserving mechanisms ensures that user data remains secure and compliant with regulations like GDPR. This holistic methodology makes the system highly effective for real-time detection of fake users and spammers, offering a robust solution for maintaining the integrity, reliability, and safety of online social networks.

Algorithm 1: Fake and Spammer Detection

Begin

1. Load dataset D
2. For each user in D:
 - a. Extract features:
 - $x_1 \leftarrow \text{Normalize}(\text{Account_Age})$
 - $x_2 \leftarrow \text{Normalize}(\text{Post_Frequency})$
 - $x_3 \leftarrow \text{Normalize}(\text{Follower_Following_Ratio})$
 - $x_4 \leftarrow \text{Normalize}(\text{Link_Post_Percentage})$
 - $x_5 \leftarrow \text{Normalize}(\text{Content_Repetition_Score})$
 - ...
 - b. Form feature vector $X = [x_1, x_2, x_3, \dots, x_n]$
3. Load trained machine learning model M
4. For each user:

```

a. Predict probability score  $P = M.predict\_proba(X)$ 
b. If  $P \geq \text{threshold}$  (e.g., 0.5):
    Label  $\leftarrow$  "Fake"
Else:
    Label  $\leftarrow$  "Genuine"

5. (Optional) Apply Explainable AI method:
    - Generate feature importance for each prediction using SHAP or LIME
    - Display which features contributed most to the "Fake" label

6. Output user ID with Label and Explanation (if enabled)

End
Begin

1. Load labeled dataset with known fake/genuine users
2. Extract and normalize features as above
3. Split dataset into Training and Testing sets
4. Train model M using classifiers like:
    - Random Forest / Decision Tree / SVM / XGBoost
5. Validate model using accuracy, precision, recall, F1-score
6. Save trained model M

End

```

The proposed algorithm for fake user and spammer detection in Online Social Networks (OSNs) operates by first extracting and normalizing a set of key features from each user profile, such as account age, posting frequency, follower-to-following ratio, percentage of posts containing links, and content repetition score. These features are combined into a numerical feature vector representing the behavioral and structural patterns of each user. A pre-trained machine learning model—such as a Random Forest, SVM, or Decision Tree—is then used to process the feature vector and calculate the probability that a user is fake. If the probability exceeds a defined threshold (e.g., 0.5), the user is labeled as "Fake"; otherwise, they are considered "Genuine." To enhance trust and transparency, the system optionally integrates explainable AI techniques (like SHAP or LIME) to identify which features most influenced the classification decision. During the offline training phase, the model is trained on a labeled dataset containing both fake and genuine profiles, allowing it to learn patterns and relationships among features. This algorithmic approach ensures real-time, interpretable, and accurate detection of fake users and spammers, while being scalable and adaptable to evolving OSN threats.

5. Simulation Environment

The simulation environment for the proposed fake user and spammer detection system is designed to replicate real-world conditions found in Online Social Networks (OSNs). It is implemented using Python and machine learning libraries such as Scikit-learn, Pandas, and NumPy, with optional support for deep learning frameworks like TensorFlow or PyTorch for advanced experimentation presented in Table 1. The dataset used for simulation consists of synthetic and real user profiles, labeled as "Fake" or "Genuine," and includes diverse feature sets such as account metadata, activity logs, and content patterns. The environment supports both training and testing phases, where the dataset is split into training (70%) and testing (30%)

subsets to evaluate model performance. Standard classification metrics like accuracy, precision, recall, and F1-score are used for evaluation. Cross-validation techniques are employed to ensure the model's robustness and generalization. Additionally, a module for explainable AI (e.g., SHAP) is integrated to provide feature importance and interpretability of model predictions. The system runs on a standard computing environment with an Intel i7 processor, 16GB RAM, and a GPU-enabled setup for deep learning tasks. This simulation environment provides a controlled, reproducible setting to test, refine, and validate the effectiveness of the proposed detection framework.

Table 1: Simulation Environment

Component	Details
Programming Language	Python
Libraries/Frameworks	Scikit-learn, Pandas, NumPy, SHAP, (Optional: TensorFlow, PyTorch)
Dataset	Real and synthetic OSN user profiles labeled as "Fake" or "Genuine"
Features Used	Account Age, Post Frequency, Follower/Following Ratio, Link Percentage, Content Repetition
Dataset Split	70% Training, 30% Testing
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score
Validation Technique	K-Fold Cross-Validation
Explainability Module	SHAP (SHapley Additive exPlanations)
System Specifications	Intel Core i7 Processor, 16 GB RAM, GPU-enabled (for deep learning tasks)
Model Types Used	Decision Tree, Random Forest, SVM, (Optional: Neural Networks)
Execution Environment	Local machine or cloud-based Python IDE (e.g., Jupyter Notebook, Colab)

6.Simulation Results and Discussion

The simulation results demonstrate the effectiveness of the proposed machine learning-based system in detecting fake users and spammers within Online Social Networks (OSNs). The models were evaluated using key performance metrics including accuracy, precision, recall, and F1-score. Among the classifiers tested, the Random Forest model achieved the highest performance, with an accuracy of approximately 96.8%, precision of 95.7%, recall of 96.2%, and an F1-score of 95.9%. The Decision Tree and Support Vector Machine (SVM) models also performed well, achieving over 90% accuracy, but with slightly lower precision and recall compared to Random Forest. These results highlight the system's ability to accurately distinguish between genuine and fake accounts based on behavioral and profile features. The use of explainable AI techniques further allowed for interpretation of the classification results, revealing that features such as link-posting frequency, account age, and content repetition had the most influence in detecting fake profiles. Additionally, the system demonstrated low false positive and false negative rates, making it reliable for practical deployment.

Table 2: Classification of ML for the Fake User Detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)	False Negative Rate (%)
Random Forest	96.8	95.7	96.2	95.9	2.1	1.8
Decision Tree	93.4	91.2	92.5	91.8	4.5	3.3

SVM (RBF Kernel)	91.7	90.1	90.6	90.3	5.2	3.9
Naïve Bayes	87.9	85.5	86.0	85.7	6.8	5.5
Logistic Regression	89.2	86.7	87.3	87.0	6.1	4.9

The performance comparison of different machine learning models for fake user and spammer detection shows that the Random Forest classifier outperforms other models across all key metrics presented in Table 2. It achieves the highest accuracy of 96.8%, indicating its strong overall ability to correctly classify both fake and genuine users. Its precision (95.7%) and recall (96.2%) values suggest that it not only accurately identifies fake users but also captures most of the actual fake accounts, leading to a balanced and reliable detection system as reflected by its F1-score of 95.9%. The low false positive rate of 2.1% and false negative rate of 1.8% further demonstrate the model's effectiveness in minimizing misclassifications. The Decision Tree model also performs well, with an accuracy of 93.4% and slightly higher error rates compared to Random Forest, while the SVM with an RBF kernel delivers respectable results but with increased false positives and negatives. Naïve Bayes and Logistic Regression models show comparatively lower performance, with accuracy below 90% and higher rates of false positives and negatives, indicating less robustness for this task.



Figure 3: Home page



Figure 4: Tweet Server Login



Figure 5: Server menu

This set of images depicts various pages of a web application focused on "Spammer Detection and Fake User Identification on Social Networks," specifically a Twitter-like platform. In figure 3 shows the Tweet Server Login Page. It features a "Welcome to Tweet Server Login Page.." heading, a prominent login button with a padlock icon, and input fields for "Userid Name (required)" and "Password (required)," along with a "Login" button to submit credentials and a "Back" link. This page clearly serves as the entry point for administrators or moderators (the "Tweet Server") to access the system's backend functionalities. Figure 4 displays the Tweet Server Page after a successful login. It welcomes the user with "Welcome to Tweet Server Page.." and presents a detailed "Server Menu" on the right sidebar. This menu offers a comprehensive set of administrative tools, including options to "View and Authorize Users,"

"Add and View Spam Filters," "View All User Posted Tweets," "View All User Tweets Based On URLs," "View Friend Request and Response," "View All Tweets with Re-Tweets," "View All Tweets, Re-Tweets and Comments," "View All Spammers Detection," "View All Fake User Identification," "View Fake User Identification Results," and "Log Out." The main content area of this page provides a textual explanation of the project's purpose, emphasizing the growing problem of spam and fake users on social networks like Twitter and the techniques used for their detection. Figure 5 presents the main page of the "Twitter - Online Social Network Service," which is likely the home or public-facing view of the application. It includes a "Sidebar Menu" with links to "Home," "Users," and "Tweet Server," suggesting navigation options for different sections of the application. Below these links, there's a "Concepts" section listing "Classification," "Fake user detection," "Online social network," and "Spammer's Identification," indicating the core technical aspects addressed by the system. Similar to the Tweet Server page, the main content area provides a detailed textual overview of the project's objectives, highlighting the prevalence of spam and fake users on social networks and the methodologies employed for their identification and detection. Collectively, these images illustrate a multi-faceted web application with a user-friendly interface for general social networking activities (implied by the "Twitter - Online Social Network Service" context) and a robust, secure backend for administrative control and the crucial function of identifying and mitigating spam and fake user accounts, thereby enhancing the integrity and reliability of the social network.



Figure 6: All posted tweets



Figure 7: view fake tweets



Figure 8: user register



Figure 9: user login



Figure 10: user menu



Figure 11: User posted tweets

Figure 6- Figure 11 illustrates the key user-facing and administrative functionalities of a "Spammer Detection and Fake User Identification on Social Networks" application, which appears to mimic a Twitter-like platform. Figure 6 displays the User Registration page,

welcoming new users and providing fields for essential information like User Name, Password, Email Address, Mobile Number, Address, Date of Birth, Gender, and the option to select a profile picture. This is the entry point for individuals to join the social network. In figure 7 shows the User Login page. This page presents input fields for "User Name" and "Password" with a "Login" button and a "New User? Register" link, guiding existing users to sign in. Once a user logs in, Figure 8 depicts the User's main page (Welcome User pavan). This page welcomes the logged-in user (e.g., "pavan") and provides a "User Menu" with options such as Home, My Profile, Search Friends, Create Tweets, View My Friends, View Friend Requests, Search Tweets and Comment, View My Tweets and Comments, View Friend's Retweets and Give Comments, and Log Out. The main content area also includes a detailed explanation of the project's goals, focusing on spammer and fake user detection, similar to previous descriptions. In figure 9 showcases the "All My Posted Tweets and It's Comments.." page, displaying a table of tweets posted by the logged-in user, including "Tweet Image," "Tweet Name," "Tweet Description," "Tweet Uses," and "Date," along with any associated comments. This represents the user's personal content feed.

In figure 10 presents the "All Posted Tweets and its Retweets." page, which provides a comprehensive view of all tweets posted on the platform, along with their associated comments and retweets. This table includes columns for "ID," "Tweet Image," "Tweet Name," "Tweet Description," "Tweet Uses," "Date," and "Tweet Owner," giving an overview of the platform's content. Figure 11 displays the "View Fake Tweet Identification Results.." page, which is likely an administrative or analytical view. It presents a bar chart titled "JS Chart" showing results related to fake tweet identification, with categories like "HP Laptop," "kids_pizza," and "Sandose_Panduva." This page highlights the system's core functionality in detecting and visualizing potentially fake or malicious content. The images demonstrate the full cycle of user interaction, from registration and login to engaging with content (posting, viewing, commenting), and also unveil the critical backend capabilities for identifying and managing problematic content like spam and fake tweets, reinforcing the application's purpose of ensuring a cleaner and more secure social networking environment.

Table 3: User Classification

User ID	Account Age (days)	Post Frequency (per day)	Link Post %	Follower/Following Ratio	Predicted Label	Actual Label	Confidence Score (%)
U001	15	12	85%	0.2	Fake	Fake	96.3
U002	450	1.5	10%	1.1	Genuine	Genuine	92.5
U003	5	20	92%	0.1	Fake	Fake	98.1
U004	320	2.3	15%	0.9	Genuine	Genuine	89.7
U005	30	8	75%	0.3	Fake	Genuine	85.2
U006	610	1.2	5%	1.3	Genuine	Genuine	94.0
U007	12	15	88%	0.2	Fake	Fake	97.5

The classification results illustrate how the proposed system analyzes various user features to accurately distinguish between fake and genuine accounts presented in Table 3. For example, user U001, with a young account age of 15 days, high posting frequency of 12 posts per day, a large proportion of posts containing links (85%), and a low follower-to-following ratio of 0.2, is correctly identified as fake with a high confidence score of 96.3%. Similarly, U003 and

U007 exhibit similar suspicious behavior patterns and are accurately classified as fake with confidence scores above 97%. In contrast, genuine users like U002, U004, and U006 have older account ages, lower posting frequencies, fewer link-based posts, and healthier follower ratios, resulting in correct genuine classifications with confidence scores ranging from 89.7% to 94%. However, there is one misclassification: user U005 is labeled as fake by the model despite being genuine, likely due to moderately suspicious features such as a relatively young account age of 30 days, high posting frequency, and significant link posting. This highlights that while the system performs well overall, borderline cases with mixed features may sometimes lead to false positives.

7. Conclusion

This paper presents an integrated machine learning framework for the effective detection of spammers and fake users in Online Social Networks (OSNs). By leveraging a comprehensive feature analysis approach combined with advanced classifiers like Random Forest, the proposed system achieves high accuracy, precision, and recall, outperforming several baseline models. The inclusion of explainable AI techniques ensures transparency in decision-making, enhancing user trust and platform accountability. Furthermore, the system's adaptability through continuous learning and its incorporation of privacy-preserving measures make it well-suited for real-time deployment in dynamic social networking environments. While the simulation results demonstrate promising performance, future work can focus on expanding the system's capabilities to handle multilingual data and more sophisticated attacker strategies. Overall, this study contributes a robust, scalable, and interpretable solution to enhance security and maintain the integrity of OSNs in the face of growing threats from fake and spam accounts.

Acknowledgement: Not Applicable.

Funding Statement: *The author(s) received no specific funding for this study.*

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

- [1] F. Siddiqui and M. Suaib, "Enhancing Spammer Fake Profile Detection on Social Media Platforms using Artificial Neural Networks," *International Journal of Engineering and Management Research*, vol.13, no.4, pp.1-6, 2023.
- [2] G. Jethava and U. P. Rao, "Exploring security and trust mechanisms in online social networks: An extensive review," *Computers & Security*, vol.140, no.103790, 2024.
- [3] F. Siddiqui and M. Suaib, "Machine learning approaches for fake user and spammer detection: A comprehensive review and future perspectives," *International Journal of Engineering and Management Research*, vol.13, no.3, pp.23-46, 2023.
- [4] K. Smita, N. Harika, N. Advaita, O. L. Kalyani and T. Kruthika, "Fake Profile Identification in Social Network Using Machine Learning and NLP," *Turkish Journal of Computer and Mathematics Education*, vol.14, no.3, pp.689-699, 2023.
- [5] S. Verma and S. Shukla, "Review of Fake Profile Classification and Identification on Social Networks," *Grenze International Journal of Engineering & Technology (GIJET)*, vol.10, 2024.
- [6] N. A. Nawaz, K. Ishaq, U. Farooq, A. Khalil, S. Rasheed, A. Abid and F. Rosdi, "A comprehensive review of security threats and solutions for the online social networks industry," *PeerJ Computer Science*, vol.9, no.e1143, 2023.
- [7] A. K. Tyagi, K. Naithani and S. Tiwari, "Security and Possible Threats in Today's Online Social Networking Platforms," *Online Social Networks in Business Frameworks*, pp.159-199, 2024.
- [8] A. Mughaid, I. Obeidat, S. AlZu'bi, E. A. Elsoud, A. Alnajjar, A. R. Alsoud and L. Abualigah, "A

-
- novel machine learning and face recognition technique for fake accounts detection system on cyber social networks,” *Multimedia Tools and Applications*, vol.82, no.17, pp.26353-26378, 2023.
- [9] S. Y. Modadugu, S. S. N. Rao and D. V. Reddy, “Extreme learning machine for spammer detection and fake user identification from twitter,” *In 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, pp.1141-1146, 2024.
- [10] M. J. Alam, I. Hossain, S. Puppala and S. Talukder, “Combating identity attacks in online social networks: A multi-layered framework using zero-knowledge proof and permissioned blockchain,” *In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp.636-643, 2023.
- [11] S. P. Jayanna, S. Venkateswarlu, B. I. Bharathi, C. H. Mahitha, P. Praharshitha and K. Nikhitha, “Fake Social Media Profile Detection And Reporting,” *Metallurgical and Materials Engineering*, pp.965-971, 2025.
- [12] A. Shah, S. Varshney and M. Mehrotra, “Threats on online social network platforms: classification, detection, and prevention techniques,” *Multimedia Tools and Applications*, pp.1-33, 2024.
- [13] Sreedhhar Bhukya, K. VinayKumar and N. C. Santosh, “A Novel Dynamic Novel Growth model for Mobile Social Networks,” *Journal of Computer Allied Intelligence*, vol.2, no.1, pp.46-53, 2024.
- [14] Pratap Rajkumar Lalwani and Varsha S. Sukhadeve, “Impact of Technology 4.0 on Quality Service, Customer Satisfaction Pertaining to E-Commerce Platforms,” *International Journal of Advance Research and Innovation*, vol.12, no.1, pp.1-6, 2024.
- [15] T. Bikku, N. S. Biyyapu, J. C. Sekhar, M. K. Kumar, S. M. Nokerov and V. K. Pratap, “The Social Network Dilemma: Safeguarding Privacy and Security in an Online Community,” *International Journal of Safety & Security Engineering*, vol.14, no.1, 2024.
- [16] E. Alnagi, A. Ahmad, Q. A. Al-Haija and A. Aref, “Unmasking Fake Social Network Accounts with Explainable Intelligence,” *International Journal of Advanced Computer Science & Applications*, vol.15, no.3, 2024.
- [17] Rekha Gangula and Rega Sravani, “Enhanced Detection of Social Bots on Online Platforms using SemiSupervised K-Means Clustering,” *Journal of Sensors, IoT & Health Sciences*, vol.2, no.1, pp.6-16, 2024.
- [18] D. Javed, N. Z. Jhanjhi, N. A. Khan, S. K. Ray, A. Al-Dhaqm and V. R. Kebande, “Identification of Spambots and Fake Followers on Social Network via Interpretable AI-Based Machine Learning,” *IEEE Access*, 2025.
- [19] A. T. Mahmood, R. K. Naser and S. K. Abd, “The Security System in Information Technology Detects Fake Accounts on the Youtube Platform,” *International Journal of Intelligent Engineering & Systems*, vol.16, no.1, 2023.
- [20] L. Caruccio, G. Cimino, S. Cirillo, D. Desiato, G. Polese and G. Tortora, “Malicious account identification in social network platforms,” *ACM Transactions on Internet Technology*, vol.23, no.4, pp.1-25, 2023.
- [21] M. Bhattacharya, S. Roy, S. Chattopadhyay, A. K. Das and S. Shetty, “A comprehensive survey on online social networks security and privacy issues: Threats, machine learning-based solutions, and open challenges,” *Security and Privacy*, vol.6, no.1, pp.e275, 2023.
- [22] W. Dracewicz and M. Sepczuk, “Detecting Fake Accounts on Social Media Portals—the X Portal Case Study,” *Electronics*, vol.13, no.13, pp.2542, 2024.
- [23] A. Podorozhniak, N. Liubchenko, V. Oliinyk and V. Roh, “Research application of the spam filtering and spammer detection algorithms on social media and messengers,” *Advanced Information Systems*, vol.7, no.3, pp.60-66, 2023.
- [24] R. Chauhan, A. Singh, C. Bhatt and H. U. Habiba, “Spammer detection and fake user detection using modified Naïve Bayes and decision Tree classifier,” *In 2023 Second International Conference on Informatics (ICI)*. IEEE, pp.1-5, 2023.
- [25] I. Haq, W. Qiu, J. Guo and T. Peng, “Spammy names detection in pashto language to prevent fake accounts creation on social media,” *In 2023 8th International Conference on Signal and Image*
-

Processing (ICSIP). IEEE, pp.614-618, 2023.

- [26] D. R. Patil, T. M. Pattewar, V. D. Punjabi and S. M. Pardeshi, "Detecting Fake Social Media Profiles Using the Majority Voting Approach," *EAI Endorsed Transactions on Scalable Information Systems*, vol.11, no.3, 2024.