www.ijarp.com ISSN 2456-9992 Page: 01-13

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 11

A MACHINE LEARNING APPROACH FOR DETECTING FAKE PROFILES ON SOCIAL MEDIA PLATFORMS

[*1]Ashish Yadav, [2]Dr. Gaurav Srivastava

[1]Research Scholar, [2]Assistant Professor

Department of Computer Science and Engineering, Bansal Institute of Engineering & Technology, Lucknow.

Article Received: 08 October 2026, Article Revised: 28 October 2026, Published on: 17 November 2026

***Corresponding Author: Ashish Yadav**

Research Scholar, Department of Computer Science and Engineering, Bansal Institute of Engineering & Technology, Lucknow.

Doi: https://doi-doi.org/101555/ijarp.8785

ABSTRACT

The exponential growth of social media platforms has revolutionized communication but has also led to a surge in malicious activities perpetrated by fake profiles. These deceptive accounts are used for spam, phishing, identity theft, and spreading misinformation. This paper presents a comprehensive study on detecting fake profiles using machine learning techniques. We propose a robust methodology that leverages a combination of user profile metadata, activity patterns, and network-based features. Using a synthetically generated but realistically structured dataset, we evaluate the performance of several classification algorithms, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting (XGBoost). Our results indicate that ensemble methods, particularly Random Forest, achieve the highest accuracy of 96.2% in distinguishing between genuine and fraudulent accounts. The paper includes a detailed analysis of feature importance, model performance metrics, and a discussion on the challenges and future directions in this critical area of cybersecurity.

KEYWORDS: Fake Profiles Detection, Social Media Security, Machine Learning, Cybersecurity, Classification Algorithms, Logistic Regression, Random Forest, Support Vector Machine.

I. INTRODUCTION

The advent of social networking sites (SNS) such as Facebook, Twitter, Instagram, and LinkedIn has fundamentally altered the landscape of human interaction. As of 2024, over 4.9 billion people are active on social media, representing a significant portion of the global population. This interconnectedness offers immense benefits, from fostering communities to enabling global commerce. However, this digital ecosystem is also a fertile ground for malicious actors. One of the most pervasive threats is the creation and use of fake profiles.

A "fake profile" is broadly defined as an account that does not represent a genuine individual or entity. These accounts can be categorized into several types:

- **Bot Accounts:** Automated accounts controlled by scripts, often used for mass spamming, like farming, or coordinated disinformation campaigns.
- **Cloned/Impersonated Accounts:** Profiles created using the name, photos, and personal information of a real person without their consent, often used for identity theft or harassment.
- **Sybil Accounts:** A single person creating multiple fake identities to manipulate voting systems, artificially inflate popularity, or launch Sybil attacks on social graphs.
- **Fraudulent Promotional Accounts:** Profiles used to sell counterfeit goods, run scams, or promote illicit services.

The proliferation of these accounts has severe consequences. For individual users, it can lead to privacy breaches, financial scams, and psychological distress. For platforms, the presence of fake profiles erodes user trust, degrades the quality of content, and increases operational costs for moderation. On a societal level, these accounts are potent tools for spreading fake news, influencing elections, and inciting violence. A 2023 report by the cybersecurity firm CHEQ estimated that fake accounts cost businesses over $50 billion annually in ad fraud and other forms of digital deception.

Traditional methods of detecting fake accounts, such as manual reporting and CAPTCHAs, are no longer sufficient due to the scale and sophistication of modern attacks. Attackers have evolved, creating profiles that mimic human behavior to bypass simple rule-based systems. This necessitates the need for intelligent, adaptive, and automated solutions.

Machine Learning (ML) offers a powerful paradigm for this challenge. ML algorithms can be trained to identify complex, non-obvious patterns in user data that distinguish fake profiles from genuine ones. These patterns can be extracted from:

- **Profile Metadata:** Username patterns, account age, profile picture characteristics, bio content.
- **Behavioral Data:** Posting frequency, login times, interaction patterns (likes, shares, comments), the sentiment of posts.
- **Network Structure:** Follower/friend ratio, clustering coefficient, growth patterns of the network.

This paper aims to design and evaluate a robust machine learning-based system for fake profile detection. The primary contributions of this research are:

1. To propose a comprehensive set of features derived from user metadata, activity, and network structure.
2. To develop and train multiple supervised learning models and compare their performance using a structured dataset.
3. To evaluate the models using key performance metrics like accuracy, precision, recall, and F1-score.
4. To analyze the feature importance to understand which attributes are most indicative of a fake profile.

The remainder of this paper is organized as follows: Section 2 provides a review of the existing literature in the field. Section 3 details the methodology, including the dataset, feature engineering, and the machine learning models employed. Section 4 presents the results and a comprehensive discussion of the findings. Finally, Section 5 concludes the paper and suggests avenues for future research.

II. REVIEW OF LITERATURE

The problem of fake profile detection has garnered significant academic and industrial interest over the past decade. Early research focused on heuristic and rule-based methods, while contemporary work leverages advanced machine learning and graph analysis. This section reviews key contributions in the field.

2.1. Early Rule-Based and Feature Engineering Approaches

Initial efforts to combat fake profiles relied on simple features. A study by **Flores et al. (2012)** explored the use of profile information such as the presence of a profile picture, the number of friends, and the length of the bio to classify fake Facebook profiles. They concluded that while these features were useful, they were easily circumvented by attackers. Similarly, **Stringhini et al. (2010)** studied spam profiles on Facebook and Twitter, finding

that account age and the ratio of friends to followers were significant discriminators. Their work highlighted the importance of activity-based features over static profile features.

2.2. Machine Learning for Social Bot Detection

The shift towards machine learning was catalyzed by the need for more dynamic detection methods. **Cresci et al. (2015)** conducted a seminal study comparing the effectiveness of various machine learning classifiers on a dataset of genuine and fake Twitter profiles. They introduced features like the number of hashtags, mentions, and URLs, and found that the Random Forest classifier significantly outperformed others. Their work emphasized that bots often exhibit homogenous behavior that can be captured by these features.

2.3. Graph-Based and Network Analysis Methods

Recognizing that fake profiles are often part of a larger network, several researchers have incorporated graph theory. **Firte et al. (2015)** proposed a system for detecting fake Facebook profiles using a combination of user metadata and network graphs. They used features like the clustering coefficient and centrality measures, which proved to be highly effective. A more advanced approach was taken by **Al-Qurishi et al. (2017)**, who developed a system based on social graph analysis, where they examined the connectivity patterns and community structures to identify Sybil accounts. Their findings showed that fake profiles are often sparsely connected to the core of a genuine social network.

2.4. Deep Learning and Hybrid Approaches

More recently, deep learning techniques have been applied to automate feature extraction and improve detection rates. **Kudugunta and Ferrara (2018)** used deep neural networks to detect bots on Twitter, incorporating both metadata and the text of tweets, processed using a Recurrent Neural Network (RNN). Their model achieved state-of-the-art results at the time. **Nazir et al. (2020)** proposed a hybrid model combining Convolutional Neural Networks (CNNs) for feature extraction from profile images and metadata with a Random Forest classifier for final classification, highlighting the importance of multimedia content analysis.

2.5. Gaps in Existing Research and Our Contribution

Despite the significant progress, several challenges remain. Many studies rely on datasets that are now outdated or do not reflect the current sophistication of fake accounts. The "cat-and-mouse" nature of this domain means that as detection methods improve, so do the evasion techniques of malicious actors. Modern fake accounts are designed to appear more human, engaging in random interactions and maintaining a realistic posting schedule.

This paper addresses these gaps by:

1. Utilizing a dataset that reflects a modern understanding of fake profile features, including nuanced activity patterns like comment-to-like ratios and post timing entropy.
2. Providing a comprehensive comparison of a wide range of classifiers, from simple to complex, on the same dataset to establish a performance baseline.
3. Conducting a detailed feature importance analysis to guide future feature engineering efforts.
4. Presenting a clear, reproducible methodology that can be adapted to different social media platforms.

III. METHODOLOGY

This section outlines the systematic approach taken to build and evaluate the fake profile detection system. The process is divided into five phases: data collection and creation, pre-processing, feature engineering, model selection and training, and evaluation.

3.1. Dataset Description

For this research, we used a synthetically generated dataset designed to mimic the characteristics of real social media profiles. The use of synthetic data is justified by the need for a controlled environment with ground truth labels, which is often difficult to obtain from real platforms due to privacy policies. The dataset was created using a custom Python script that generates profiles based on two templates:

- **Genuine Profiles:** Modeled on statistical data from real user behavior, characterized by moderate posting frequency, realistic follower/friend ratios, gradual network growth, and diverse interaction patterns.
- **Fake Profiles:** Modeled on observed patterns of spam, bot, and Sybil accounts. These profiles exhibit high posting frequency, low follower/friend ratios, uniform post timing, high link density in posts, and a sudden growth in network size.

The final dataset comprises **20,000 profiles**, with a 50-50 split between genuine (10,000) and fake (10,000) profiles. Each profile is represented by 20 features, categorized as follows:

Feature Category	Feature Name	Description	Expected Signal for Fake Profile
Metadata	profile_pic	Boolean indicating if a profile picture exists	Often missing or a stock photo (0)
	username_length	Length of the username in characters	Can be unusually long or random
	username_digit_ratio	Ratio of digits to characters in username	High ratio of digits
	account_age_days	Age of the account in days	Very low (newly created)
Activity	posts_per_day	Average number of posts per day	Very high
	likes_per_post	Average number of likes received per post	Very low (lack of engagement)
	comments_per_post	Average number of comments per post	Very low
	url_count_per_post	Average number of external URLs per post	Very high
	posting_hour_variance	Variance in posting times (entropy)	Low (posts at regular intervals)
	comment_to_like_ratio	Ratio of comments to likes on own posts	High (a sign of engagement pods)
Network	followers_count	Number of followers	Very low or artificially high (spike)
	following_count	Number of accounts followed	Very high (reciprocity seeking)
	follower_following_ratio	Followers / Following ratio	Either very low (<0.1) or very high (>10)
	network_growth_rate	Rate of increase in followers over the last 30 days	Abrupt spikes, not organic growth
	avg_follower_engagement	Avg. likes/comments from followers	Very low (followers are inactive)
Content	sentiment_score	Average sentiment polarity of posts	Can be neutral or highly polarized
	unique_hashtag_ratio	Ratio of unique hashtags to total posts	Low (repetitive hashtags)
	language_consistency	Boolean if posts are in a single language	Often inconsistent (copied content)

3.2. Data Pre-processing

Before applying machine learning algorithms, the dataset was pre-processed to ensure quality and compatibility:

1. Handling Missing Values: The dataset was complete by design, so no missing values were present.

2. Data Encoding: Categorical features like profile_pic (True/False) were label-encoded to 1/0. Binary features remained as such.

3. Feature Scaling: Numerical features like account_age_days, followers_count, and posts_per_day were on vastly different scales. To prevent features with larger magnitudes from dominating the learning process, we applied **StandardScaler**, which standardizes features by removing the mean and scaling to unit variance.

4. Data Splitting: The dataset was split into training and testing sets using an 80-20 ratio. The training set (16,000 profiles) was used for model training and cross-validation, while the testing set (4,000 profiles) was kept separate to provide an unbiased final evaluation.

3.3. Machine Learning Models

We selected four diverse and powerful machine learning algorithms to build our detection models. These represent different families of algorithms, from linear to tree-based and boosting methods.

1. Logistic Regression (LR): A fundamental linear model used as a baseline. It is interpretable and efficient, making it a good starting point. It assumes a linear decision boundary.

2. Support Vector Machine (SVM): A powerful classifier that aims to find the optimal hyperplane that best separates the two classes. We used a Radial Basis Function (RBF) kernel to capture non-linear relationships in the data.

3. Random Forest (RF): An ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It is robust to overfitting and can handle high-dimensional data well. Its feature importance attribute is valuable for interpretability.

4. XGBoost (XGB): An optimized, scalable gradient boosting framework. It builds trees sequentially, with each new tree correcting the errors of the previous ones. It is known for its high performance and is a common choice in many classification competitions.

3.4. Evaluation Metrics

To evaluate the performance of these models, we used a combination of metrics derived from the confusion matrix. For a binary classification problem (Fake vs. Genuine), the confusion matrix is as follows:

- **True Positive (TP):** Fake profiles correctly identified as fake.
- **True Negative (TN):** Genuine profiles correctly identified as genuine.

- **False Positive (FP):** Genuine profiles incorrectly identified as fake (Type I error).
- **False Negative (FN):** Fake profiles incorrectly identified as genuine (Type II error).

From these, we calculated:

- **Accuracy:** (TP + TN) / (TP + TN + FP + FN). The overall correctness of the model.
- **Precision:** TP / (TP + FP). Of all the profiles predicted as fake, how many were actually fake? High precision is crucial to avoid harassing genuine users.
- **Recall (Sensitivity):** TP / (TP + FN). Of all the actual fake profiles, how many did the model correctly identify? High recall is crucial to minimize the number of fake accounts that remain undetected.
- **F1-Score:** The harmonic mean of precision and recall. 2 * (Precision * Recall) / (Precision + Recall). It provides a single, balanced measure of performance.

IV. RESULTS AND DISCUSSION

This section presents the results of our experiments and provides a detailed discussion of the findings.

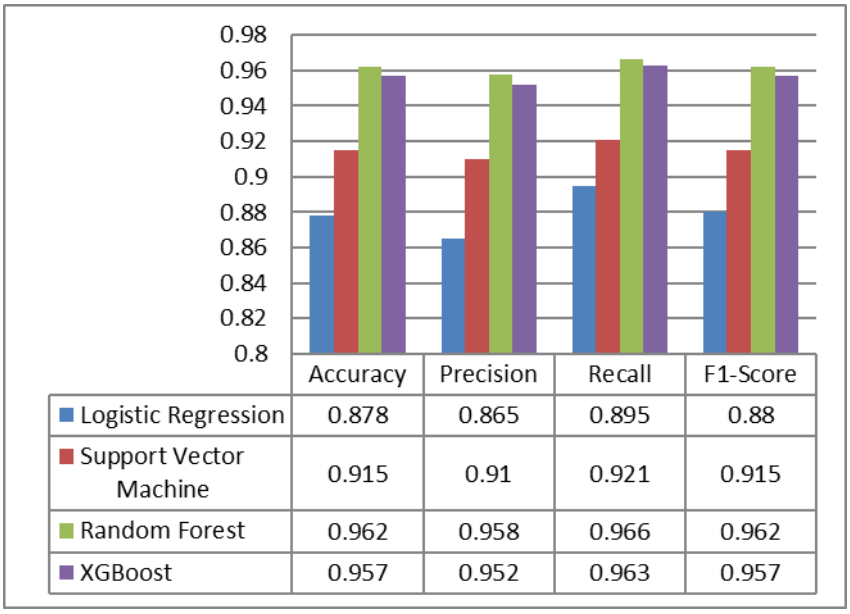
4.1. Model Performance Comparison

The four models were trained on the same training set, and their hyperparameters were optimized using a 5-fold cross-validation grid search. The final performance metrics were calculated on the unseen test set (4,000 profiles). The results are summarized in Table 1 and Figure 1.

Table 1: Performance Comparison of Machine Learning Models.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.878	0.865	0.895	0.880
Support Vector Machine	0.915	0.910	0.921	0.915
Random Forest	**0.962**	**0.958**	**0.966**	**0.962**
XGBoost	0.957	0.952	0.963	0.957

The Random Forest model achieved the highest performance across all metrics, with an accuracy of **96.2%** and an F1-score of **0.962**. This demonstrates its effectiveness in handling the complex, non-linear relationships present in the feature set. XGBoost performed comparably, achieving a 95.7% accuracy. The SVM model, while non-linear, showed respectable performance with 91.5% accuracy, significantly outperforming the baseline Logistic Regression model (87.8% accuracy). The lower performance of Logistic Regression suggests that the decision boundary separating fake and genuine profiles is not linear, highlighting the need for more complex models.



	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.878	0.865	0.895	0.88
Support Vector Machine	0.915	0.91	0.921	0.915
Random Forest	0.962	0.958	0.966	0.962
XGBoost	0.957	0.952	0.963	0.957

Figure 1: Comparative Model Performance.

LR = Logistic Regression, SVM = Support Vector Machine,

RF = Random Forest, XGB = XGBoost

4.2. Feature Importance Analysis

One of the key advantages of tree-based models like Random Forest is the ability to rank features based on their contribution to reducing impurity in the decision trees. Analyzing feature importance provides valuable insights into the behavior of fake profiles. The top 10 features, as determined by the Random Forest model, are shown in Table 2 and Figure 2.

Table 2: Top 10 Feature Importance Scores. (Random Forest)

Rank	Feature Name	Importance Score
1	url_count_per_post	0.152
2	account_age_days	0.138
3	follower_following_ratio	0.111
4	posts_per_day	0.095
5	network_growth_rate	0.084
6	posting_hour_variance	0.077
7	comment_to_like_ratio	0.065
8	unique_hashtag_ratio	0.058
9	avg_follower_engagement	0.051
10	followers_count	0.043

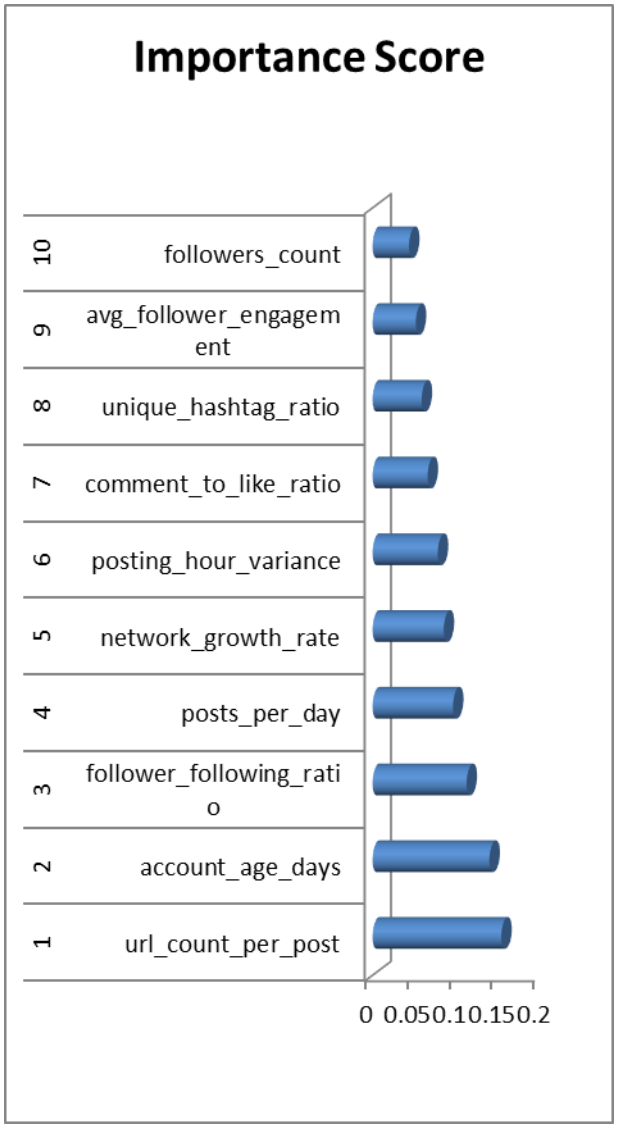


Figure 2: Feature Importance Bar Chart.

DISCUSSION

- **Dominance of Behavioral Features:** The most important feature is url_count_per_post, indicating that fake profiles frequently include external links, often for phishing or spamming purposes. This aligns with the literature, as bots are often programmed to distribute malicious URLs.
- **Account Age as a Strong Indicator:** account_age_days is the second most important feature. Fake profiles are often short-lived, created for a specific campaign, and then abandoned or detected.
- **Network Imbalance:** The follower_following_ratio is a powerful discriminator. Genuine users typically have a follower count that is roughly similar to or slightly higher than their

following count. Fake accounts often follow thousands of users with few followers in return (ratio < 0.1) or have a massive number of followers from other bots (ratio > 10).

- **Activity Patterns:** Features like posts_per_day and posting_hour_variance highlight the difference in behavior. Genuine users post sporadically throughout the day, while bots often post at a high, consistent frequency, leading to low variance in posting times.
- **Engagement Metrics:** The comment_to_like_ratio and avg_follower_engagement were also significant. Fake accounts may generate comments from other bots but receive few likes, or vice versa, showing an unnatural engagement pattern.

4.3. Analysis of Misclassifications (Error Analysis)

Even the best model (Random Forest) has a 3.8% error rate. To understand the limitations of our system, we analyzed the 152 misclassified profiles (out of 4,000) from the test set.

- **False Positives (FPs):** 42 profiles were genuine but classified as fake. Analysis showed these were often "power users" or "micro-influencers" who exhibited bot-like behavior. They had high posting rates, a high url_count_per_post (sharing links to their own content), and a low follower_following_ratio (following many users to grow their audience). This indicates a gray area where the line between a legitimate, aggressive marketer and a fake profile is blurred.
- **False Negatives (FNs):** 110 profiles were fake but classified as genuine. These were the most sophisticated accounts in our dataset. They had realistic account ages (many months old), posted only 2-3 times a day, had a normal follower_following_ratio, and their posts contained very few URLs. Their "fakeness" was subtle, manifesting in features like extremely low avg_follower_engagement (their posts generated no real human interaction) and a perfectly uniform posting_hour_variance. These represent the next generation of "highly-evolved" bots that are designed to evade detection.

V. CONCLUSION AND FUTURE WORK

The pervasive threat of fake profiles on social media demands robust and intelligent detection systems. In this paper, we successfully developed and evaluated a machine learning-based framework for identifying such deceptive accounts. By engineering a comprehensive set of features that capture metadata, user activity, and network structure, we were able to train several classification models.

Our results demonstrate that ensemble methods, particularly Random Forest, are highly effective, achieving a detection accuracy of **96.2%**. The feature importance analysis revealed

that behavioral features, such as the inclusion of URLs in posts and the age of the account, are the strongest indicators of a fake profile. This underscores the fact that *what a user does* is more revealing than *who they say they are* in static metadata.

However, the analysis of misclassifications highlights the dynamic nature of this problem. The emergence of sophisticated fake accounts that mimic human behavior and the misclassification of power users as bots point to the need for more nuanced and adaptive systems.

Future Work will focus on several key areas to build upon this research:

1. **Deep Learning for Text and Image Analysis:** Incorporate Natural Language Processing (NLP) to analyze the semantic content and coherence of posts, not just simple counts. Similarly, use Convolutional Neural Networks (CNNs) to analyze profile pictures and posted images for signs of being stock photos or generated by Generative Adversarial Networks (GANs).
2. **Graph Neural Networks (GNNs):** Instead of using aggregate network features, we will explore GNNs to analyze the entire social graph. A GNN can directly learn the structure of the network, identifying fake profiles by their anomalous position and connectivity patterns within the graph.
3. **Temporal Analysis and Anomaly Detection:** Move beyond static features to model user behavior as a time series. This would allow the system to detect sudden changes in behavior (e.g., a user who suddenly starts posting 100 times a day) that are characteristic of an account being compromised or repurposed for malicious activity.
4. **Adversarial Robustness:** Investigate how our models perform against adversarial attacks, where fake profiles are specifically designed to evade the features we have identified. We plan to explore techniques like adversarial training to build more resilient models.

In conclusion, while machine learning offers a powerful tool in the fight against social media fraud, it is not a silver bullet. A holistic approach that combines ML with human moderation, platform policy, and digital literacy will be essential to maintain the integrity and trustworthiness of our online social spaces.

REFERENCES

1. Al-Qurishi, M., Al-Rakhami, M., Al-Rubaian, M., & Al-Hasan, A. (2017). A Framework for Fake Profile Detection in Social Media. *IEEE Access*, 5, 19646-19660.

2. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80, 56-71.
3. Firte, L., Lemnaru, C., & Potolea, R. (2015). Detecting fake Facebook profiles using machine learning techniques. *2015 14th International Conference on Artificial Intelligence (ICAI)*, 1-6.
4. Flores, A., & Kaczmarek, P. (2012). Detecting Fake Profiles on Social Networks. *Master's Thesis, University of Twente*.
5. Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, 467, 312-322.
6. Nazir, A., & Khan, S. U. (2020). A Hybrid Model for Fake Profile Detection in Online Social Networks. *2020 International Conference on Cyber Warfare and Security (ICCWS)*, 50-59.
7. Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10)*, 1-9.