International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Identification of Spammers and Fake Users in Social Networks Using Machine Learning

Harini S[1], Prof. Seema Nagaraj[2]

[1] *Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India (1BI23MC042)*
[2]*Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India*

Abstract

The integrity of online social networks is seriously threatened by spammers and phony users, especially on sites like Instagram where user-generated content and interactions are vital in determining engagement. Through the analysis of behavioural, content-based, and network-driven features, this project suggests a machine learning-based method for the automated detection of spammers and phony accounts. A supervised learning model is created and trained to accurately differentiate between spam accounts and real users. Through an easy-to-use interface, users and administrators can upload datasets or keep an eye on user activity thanks to the trained model's integration into a web application developed with the Django framework. The system facilitates early detection of malicious accounts and helps to maintain a trustworthy online environment by offering instant classification results. This work shows how web technologies and machine learning can be combined to create scalable, useful, and easily accessible social network security tools.
Automated detection systems are essential for protecting user experience as social media platforms are increasingly being abused for malicious activities, spam, and scams. In user behaviour analysis and anomaly detection, machine learning algorithms—in particular, ensemble approaches and classification models—have demonstrated encouraging results. But manually keeping an eye on millions of accounts is unfeasible and prone to errors. The suggested system reduces human error, speeds up the detection of fraudulent accounts, and builds trust in online communities by automating the detection process.

Key Words: Django-based web application, machine learning classifier, Instagram spam identification, fake user detection, Python-driven analysis, Scikit-learn models, Matplotlib, User Behaviour Analytics, Pandas-Powered Processing, Bootstrap, and HTML/CSS Interface for Data Visualization.

1. INTRODUCTION

In today's social networks, spammers and phony users are becoming a bigger problem, endangering community integrity, platform security, and user trust. With millions of daily active users, platforms like Instagram are especially susceptible to these types of malicious activity. False accounts are frequently used to propagate false information, launch phishing scams, or inflate engagement metrics, all of which have a detrimental effect on users and companies. Early detection of these accounts is essential to preserving a trustworthy and secure social networking environment.

Conventional detection techniques mainly rely on rule-based systems or manual monitoring, which are laborious, ineffective, and frequently unable to keep up with spammers' changing tactics. This difficulty emphasizes the requirement for automated, intelligent systems that can precisely spot questionable user behaviour in real time. New possibilities for social media security have been made possible by recent developments in machine learning and artificial intelligence. Machine learning models can accurately distinguish between spammers and real users by utilizing behavioural characteristics, activity patterns, and interaction metrics.

The need to create a scalable, effective, and useful method for identifying fraudulent Instagram users is what spurred this project. With little assistance from humans, the system uses supervised machine learning techniques to evaluate user activity data and categorize accounts as authentic or fraudulent. By visualizing user behaviour patterns, the solution is further improved and provides a deeper understanding of how spammers function.

The model is implemented via a web application based on the Django framework to guarantee usability and accessibility. Without requiring extensive technical knowledge, this platform allows administrators or researchers to upload datasets, visualize analytical results, and categorize accounts through an intuitive interface. To offer dependable functionality and an engaging front-end experience, technologies like Python, Scikit-learn, Pandas, Matplotlib, Bootstrap, and HTML/CSS are combined.

Designing and implementing an automated spammer and fake user detection system using machine learning techniques is the main goal of this research. Additionally, a useful web-based interface for real-world applications will be provided. Data pre-processing, feature extraction, model training, assessment, and deployment within a Django environment are all included in the scope.

This study intends to support ongoing efforts to make social media platforms more secure, reliable, and resistant to spam and fraudulent activity by fusing machine learning with web technologies.

2. LITERATURE SURVEY

The automatic detection of spammers and fraudulent users in social networks through machine learning has been the subject of a great deal of research. To increase accuracy and dependability, a large number of these studies suggest better feature selection, behavioural analysis, and classification techniques. Thanks to the availability of annotated datasets and advancements in learning algorithms, machine learning has made significant strides in the detection of fake Instagram users in recent years.

A. Detecting Spam Comments On Indonesia's Instagram Posts [1]

This study [1] suggests using machine learning to detect spam comments on Indonesian Instagram posts. The authors developed classification models that can differentiate between spam and genuine comments by extracting and evaluating textual features. The study shows how automated detection methods can enhance social media moderation.

B. Enhancing Online Security: Detection Of Fake Profiles On Instagram Using GBM [2]

The writers in [2] suggests a GBM-based model that uses network behaviour, activity, and engagement to identify phony Instagram profiles. The model's high accuracy demonstrated how well GBM works to increase social network security.

C. Advanced Algorithmic Approaches For Scam Profile Detection On Instagram [3]

The authors of [3] examine sophisticated algorithms for spotting fraudulent Instagram profiles, focusing on network-based and behavioural characteristics. Their method strengthened platform security and improved detection accuracy. Instagram Fake and Automated Account Detection [4]

D. Instagram Fake And Automated Account Detection [4]

The work in [4] presents methods for detecting automated and phony Instagram accounts. The method successfully distinguishes real users from bots by examining user profiles, activity trends, and interaction patterns. The study emphasizes how automated detection can increase platform security and trust.

E. Instagram Fake Profile Detection Using Machine Learning [5]

The study in [5] uses machine learning methods to identify phony Instagram profiles by looking at user activity, engagement patterns, and profile attributes. Their model successfully separates legitimate users from accounts that are fraudulent. The results highlight how crucial ML is to improving social media security.

F. Instagram Fake Profile Section – A Review [6]

The study in [6] offers an overview of current methods for identifying phony Instagram profiles. It compares the efficacy of network-based, behavioral, and machine learning approaches. This survey identifies present issues and makes recommendations for future steps to increase detection accuracy.

G. Instagram spam detection (ISD) [7]

The authors of [7] describe an Instagram Spam Detection system that detects and filters spam accounts and content. The method depends on examining posting patterns, user activity, and interaction trends. By lowering spam and enhancing the platform's user experience, the suggested system increases dependability.

H. **Automated Detection Of Spam On Instagram Using Classifiers And NLP [8]** Researchers in [8] propose a system that uses machine learning classifiers and natural language processing to detect spam on Instagram. By looking at textual content and

engagement patterns, the method improves the accuracy of identifying malicious activity. This work emphasizes the value of NLP in improving social media security.

I. Improved Instagram Spam Detection With Convolutional Attention Networks And Water Wave Optimization [9]

The method in [9] improves Instagram spam detection by combining water wave optimization with convolutional attention networks. The technique improves feature extraction and optimization, which increases the accuracy of separating spam from legitimate content. The potential of hybrid deep learning and optimization techniques for social media security is demonstrated by this study.

J. Detecting Fake Accounts On Instagram Using Machine Learning And Hybrid Optimization Algorithms [10]

The authors of [10] suggest using hybrid optimization algorithms in conjunction with a machine learning model to identify phony Instagram accounts. To increase classification accuracy, the system assesses user behaviour, profile characteristics, and engagement trends. The study emphasizes how well ML and optimization work together to detect fake accounts.

III. METHODOLOGY

The goal of this study's research methodology is to use machine learning techniques to create a precise and effective system for identifying spam and phony Instagram profiles. Data collection, data pre-processing, feature extraction, model design, training and evaluation, and deployment are the sequential steps that make up the entire process.

1. Information Gathering

In addition to manually annotated profiles classified as real or fake/spam, the data is gathered from publicly accessible Instagram datasets. Features like user activity patterns, content engagement, network connections, and profile details (username, bio, and number of followers) are all included in the datasets.

2. Pre-processing Data

To increase accuracy and guarantee consistency, pre-processing is done before the data is fed into the model. Categorical variables are encoded, missing values are handled, and superfluous attributes are eliminated. A predetermined range is used to normalize numerical features like followers, following ratio, and post frequency. Stop words, special characters, and tokenization are used to clean up text-based features like bios and comments.

3. Extraction of Features

For classification, the study extracts features based on users, content, and networks. Follower-to-following ratio, post count, and account age are examples of user-based features. While network-based features examine connectivity and interaction patterns, content-based features address sentiment in captions, hashtag usage, and comment style. The model can differentiate real accounts from spam or fraudulent ones thanks to these combined features.

4. Design of the Model

The supervised machine learning algorithms Random Forest, Logistic Regression, SVM, and Gradient Boosting Machine (GBM) are assessed. Because of its capacity to manage unbalanced datasets and capture intricate relationships, GBM is selected as the main model. The model produces a binary classification genuine or fake/spam after being trained with extracted features. Performance is maximized through hyper parameter tuning.

5. Training and Evaluation

The dataset is divided into training, validation, and test sets. Training data is used to build the model, validation data helps tune hyper parameters, and the test set evaluates final performance. The model is optimized using an appropriate loss function such as log-loss. Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC to ensure balanced performance in detecting fake accounts. A confusion matrix is used to analyse misclassification patterns.

6. Implementation

Following satisfactory performance, Django is used to deploy the trained model into an online application. Users can enter information about their Instagram accounts (such as profile details or features that have been extracted) and get real-time predictions about whether the account is legitimate or spam. The web application processes uploaded data, generates predictions in real-time, and displays the results with an option to analyze another account. It does this by following Django's Model-View-Template (MVT) pattern.

Further Information On The Methodology

To improve robustness, the machine learning pipeline uses ensemble learning, cross-validation, and feature scaling. Techniques like class weighting and SMOTE (Synthetic Minority Oversampling Technique) are used to address class imbalance, where there are more real profiles than fake ones. A learning rate of 0.05, a maximum depth of 5, early stopping, and regularization to enhance generalization are the optimal parameters for the GBM model. A Django-based web application is used for deployment, processing user inputs in real-time to produce predictions instantly and offering researchers, social media analysts, and security specialists an easily navigable tool.

IV. RESULT

Accuracy, precision, recall, and F1-score are common classification metrics that were used to assess the performance of the suggested system. The system produced the following outcomes on the test dataset: The system's high recall score lowers the possibility of false negatives by successfully identifying the majority of spammers and phony accounts. Additionally, the model exhibits a good balance between sensitivity and precision, both of which are necessary for trustworthy social media monitoring. The suggested system provides better accuracy and faster inference time than some current machine learning techniques and conventional rule-based methods. Because of this, it can be used to monitor and identify spammers and phony Instagram accounts in real time.

Metric	Score (%)
Accuracy	93.4
Precision	91.2
Recall	94.7
F1-Score	92.9

Table 1: Performance Table

The model's high recall value suggests that it is very good at spotting spam and fake accounts, which is important for social media security because malicious users who go unnoticed can compromise the integrity of the platform. Because of sporadic false positives, which can cause real users to be incorrectly classified, precision is somewhat reduced; however, this trade-off is acceptable in security applications where identifying suspicious profiles is more crucial. Our model is suitable for practical deployment on large-scale social media platforms because it maintains efficiency while achieving competitive performance when compared to related works in the literature.

Figure 1: Performance Metrics

The accuracy, precision, recall, and F1-score attained by the suggested machine learning model (GBM) for Instagram fake profile detection on the test dataset are displayed in the chart.

V. CHALLENGES AND LIMITATIONS

Even though AI and machine learning techniques have the potential to identify spammers and fraudulent users in social networks, the system's development was fraught with a number of difficulties and restrictions:

A. Data Availability and Quality:

It can be difficult to find a sizable, varied, and labelled dataset of both real and fraudulent accounts. The datasets that are available might not accurately reflect real-world situations, and many social networks have privacy restrictions.

B. Changing Strategies for Spam:

Spammers constantly modify their tactics to evade detection systems. Without frequent updates and retraining, this dynamic behaviour makes it challenging to maintain a model that performs well over time.

C. Unbalanced Data:

Usually, there are a lot fewer fake accounts than real users, which leads to problems with class imbalance. This may result in biased models that have trouble identifying infrequently occurring fraudulent accounts.

D. Feature Selection and Representation:

It can be difficult to determine which features—like network connections, account activity patterns, or textual content—are most pertinent. Features that are noisy or irrelevant may make the system less accurate.

E. Computational Complexity:

Scalability may be constrained by the high computational costs associated with training and real-time detection for some sophisticated machine learning models, such as deep learning-based graph or network analysis models.

F. False Positives and Negatives:

The system may fail to identify some phony accounts (false negatives) or mistakenly flag legitimate users as fraudulent (false positives), even with high accuracy. It's still very difficult to strike a balance between recall and precision.

G. Generalization Across Platforms:

Because user behaviour, platform features, and spam tactics vary across social networks, models trained on one may not function well on another.

VI. CONCLUSION

Using machine learning and artificial intelligence techniques, this study demonstrated an automated system for detecting Instagram spammers and fraudulent users. The system demonstrated encouraging accuracy, precision, and recall by utilizing feature engineering, classification algorithms, and, when appropriate, deep learning techniques. Preprocessing, data balancing, and network-based feature extraction were used to increase the model's resilience, and its incorporation into a scalable application framework guarantees administrators and moderators on Instagram can use it practically. All things considered, the system shows how AI-driven solutions can improve the identification and mitigation of phony accounts, making the platform safer and more reliable.

According to the findings, AI-powered solutions can greatly cut down on the time and effort needed to identify phony or spam accounts, helping Instagram administrators keep their network safe and dependable. Because of the model's high recall, the majority

of phony accounts are accurately recognized, which is essential for stopping the spread of malicious activity, scams, and false information.

The current system's architecture can be modified to detect other malicious behaviors, like phishing attempts or bot activity, even though its primary focus is on spam and fake user detection. Additionally, the deployment strategy guarantees that the system can be updated and scaled with little technical assistance, which makes it feasible for a big, dynamic platform like Instagram.

All things considered, this project shows how AI and machine learning can improve Instagram's security and user confidence. It emphasizes how crucial it is to combine realistic deployment techniques with model accuracy for real-world applicability. The system may help identify malicious and fraudulent accounts more quickly, accurately, and scalablely with further development and modification, making Instagram a safer and more dependable place in the long run.

VII. REFERENCES

[1] A. A. Septiandri and O. Wibisono, "Detecting Spam Comments on Indonesia's Instagram Posts," *Journal of Physics: Conference Series*, vol. 801, no. 1, p. 012069, 2017.

[2] A. Verma, A. Warsi, and A. Kumar, "Enhancing Online Security: Detection of Fake Profiles on Instagram Using GBM," *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, vol. 12, no. 2, pp. 234–240, 2025.

[3] M. Al-Qurishi, A. Al-Rubaian, and S. A. Hameed, "Advanced Algorithmic Approaches for Scam Profile Detection on Instagram," *MDPI Electronics*, vol. 12, no. 8, p. 1571, 2023.

[4] F. C. Akyon and E. Kalfaoglu, "Instagram Fake and Automated Account Detection," *arXiv preprint* arXiv:1910.03090, 2019.

[5] S. Gharte, A. Rajguru, and S. Khamkar, "Instagram Fake Profile Detection Using Machine Learning," *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, vol. 12, no. 1, pp. 135–142, 2025.

[6] A. Yadav, R. Kumari, and R. Sharma, "Instagram Fake Profile Detection – A Review," School of Computer Science and Engineering, Greater Noida, India, 2025.

[7] P. V. Dhote, R. Ramteke, A. R. Patel, P. Dewalker, and A. Chaube, "Instagram Spam Detection (ISD)," *International Journal of Trend in Scientific Research and Development (IJTSRD)*, vol. 8, no. 5, pp. 573–578, Sep.–Oct. 2024.

[8] K. Priyadarsini, D. Chatter, and A. Dhand, "Automated Detection of Spam on Instagram using Classifiers and NLP," in *Proc. International Conference on Recent Trends in Data Science and its Applications*, Chennai, India, 2024, p. 664, doi: rp9788770040723.127.

[9] R. N., T. Thomas, and R. Agnihotri, "Improved Instagram Spam Detection with Convolutional Attention Networks and Water Wave Optimization," *Journal of Information Systems Engineering and Management*, vol. 10, no. 25s, 2025. e-ISSN: 2468-4376.

[10] P. Azami and K. Passi, "Detecting Fake Accounts on Instagram Using Machine Learning and Hybrid Optimization Algorithms," *Algorithms*, vol. 17, no. 10, p. 425, 2024.