Detection of Fake Instagram Accounts and Spam Comments Using Machine Learning Techniques

Sumedha Mahadev Sawant
Department of MSc Information Technology
D. G. Ruparel College of Arts, Science and Commerce
Mumbai, India

---***---

Abstract - Social media platforms such as Instagram have become highly vulnerable to fake accounts and spam comments, which negatively impact user trust, engagement, and platform security. Manual moderation techniques are inefficient and unable to scale with the growing volume of content. This research paper presents InstaGuard, a dual machine learning–based system designed to detect fake Instagram accounts and spam comments in real time. The proposed system utilizes a Random Forest classifier for account authenticity detection based on profile metadata and a Logistic Regression model combined with TF-IDF vectorization for spam comment classification. Feature engineering techniques such as follower–following ratio, bio characteristics, username patterns, and textual n-grams are applied to enhance prediction accuracy. The system is implemented using a Flask-based backend integrated with a web dashboard that provides confidence scores and explainable outputs. Experimental evaluation demonstrates that the proposed approach effectively identifies suspicious accounts and spam comments with high reliability. The results indicate that InstaGuard offers a practical, interpretable, and scalable solution for improving safety on social media platforms.

Key Words: Instagram Spam Detection, Fake Account Detection, Machine Learning, Natural Language Processing, TF-IDF, Random Forest, Logistic Regression, Social Media Security

1.INTRODUCTION

Social media platforms have become an integral part of modern communication, enabling users to share content, interact, and build online communities. Among these platforms, Instagram has gained massive popularity due to its visual-centric features and ease of engagement. However, the rapid growth of Instagram has also led to an increase in malicious activities such as fake accounts and spam comments [2], [7]. These activities negatively affect user trust, reduce content quality, and may lead to cyber fraud, misinformation, and harassment.

Fake Instagram accounts are commonly created to spread promotional content, execute scams, manipulate follower counts, or automate engagement using bots [1]. Similarly, spam comments often contain irrelevant advertisements, phishing links, or repetitive content that disrupts meaningful interactions on posts [3]. Traditional manual moderation and keyword-based filtering techniques are insufficient to handle the scale and evolving nature of such malicious behavior [4].

Machine learning and natural language processing techniques have shown promising results in detecting abnormal patterns in user behavior and textual data. By analyzing account metadata and comment content, intelligent systems can automatically classify accounts and comments as genuine or suspicious [2], [8]. However, many existing solutions focus only on spam comment detection or fake account identification independently and lack real-time deployment and interpretability.

This research presents InstaGuard, a machine learning-based system designed to detect both fake Instagram accounts and spam comments in an integrated manner. The proposed system employs supervised learning models to analyze account-level features and textual patterns in comments. Additionally, the system is implemented as a real-time web application using a Flask backend and a user-friendly dashboard, providing confidence scores and explainable predictions. The primary objective of this work is to develop an effective, interpretable, and scalable solution that enhances safety and reliability on Instagram.

2. LITERATURE REVIEW

The problem of detecting fake accounts and spam comments on social media platforms has attracted significant research attention in recent years. Various machine learning and natural language processing techniques have been proposed to address the growing challenges of online spam and malicious activities, particularly on platforms such as Instagram and Twitter [1], [2].

Several studies have focused on spam comment detection using traditional machine learning approaches. Techniques such as Naïve Bayes, Support Vector Machines (SVM), Random Forest, and Logistic Regression have been widely applied in combination with text preprocessing and feature extraction methods like Bag-of-Words and TF-IDF [3], [4]. These approaches have demonstrated satisfactory performance in identifying spam comments based on textual patterns, keywords, and frequency-based features. However, most of these systems operate offline and lack real-time deployment capabilities.

Recent research has also explored deep learning models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and transformer-based architectures for spam detection [6], [7]. While these models often achieve higher accuracy, they require large datasets, high computational resources, and longer training times. Moreover, deep learning models typically act as black boxes, offering limited explainability, which can reduce trust in automated moderation systems.

In addition to comment-level analysis, some researchers have investigated fake account detection by analyzing user profile metadata such as follower–following ratios, account age, posting behavior, and profile completeness [1], [4]. Machine learning classifiers like Random Forest and Gradient Boosting have shown effectiveness in identifying bot-like and suspicious accounts. Nevertheless, many existing solutions focus solely on account authenticity and do not integrate comment-level spam detection.

A key limitation observed in the literature is that most studies address either spam comments or fake accounts as isolated problems. Furthermore, few systems provide a complete end-to-end implementation that includes real-time prediction, user interaction, confidence scoring, and interpretability [7], [8]. These gaps highlight the need for a unified system that can detect both fake accounts and spam comments while offering transparency and practical deployment.

The proposed InstaGuard system aims to bridge these gaps by combining account authenticity detection and spam comment classification within a single framework. By integrating machine learning models with a real-time web-based interface and explainable outputs, this research extends existing work toward a more practical and scalable solution for social media safety.

3. PROPOSED SYSTEM ARCHITECTURE

The proposed InstaGuard system is designed as a dual-module architecture that performs fake account detection and spam comment detection in an integrated manner. The system follows a client–server model and consists of data input, feature extraction, machine learning classification, and result visualization components [2], [7].

3.1 SYSTEM OVERVIEW

The system accepts user input in the form of an Instagram username, profile URL, or comment text through a web-based dashboard. The input is processed by a Flask-based backend server, which coordinates data extraction, feature engineering, and prediction using pre-trained machine learning models. The final results are displayed to the user along with confidence scores and explanatory insights.

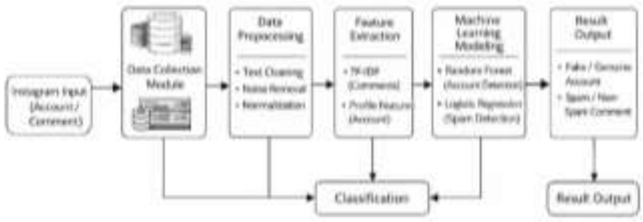
Instagram Input (Account / Comment)
Data Collection Module
Data Preprocessing
• Text Cleaning
• Noise Removal
• Normalization
Feature Extraction
• TF-IDF (Comments)
• Profile Features (Account)
Machine Learning Modeling
• Random Forest (Account Detection)
• Logistic Regression (Spam Detection)
Result Output
• Fake / Genuine Account
• Spam / Non-Spam Comment
Classification
Result Output

Fig. 1: Proposed InstaGuard System Architecture

Fig. 1: Proposed InstaGuard system architecture illustrating data collection, preprocessing, feature extraction, machine learning classification, and result visualization for detecting fake Instagram accounts and spam comments.
Figure 1 illustrates the overall system architecture of the proposed InstaGuard framework.

3.2 FAKE ACCOUNT DETECTION MODULE

The fake account detection module analyzes Instagram profile metadata to identify suspicious or bot-like accounts. Features such as number of followers, number of accounts followed, total posts, bio length, username patterns, and account visibility are extracted and processed. These features are fed into a Random Forest classifier, which predicts whether the account is genuine or suspicious. Random Forest is chosen due to its robustness, ability to handle non-linear relationships, and interpretability.

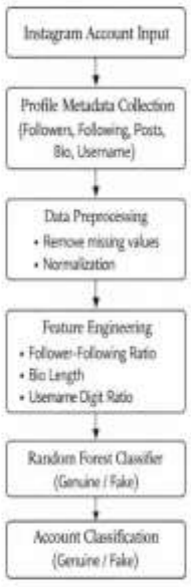
Instagram Account Input
Profile Metadata Collection (Followers, Following, Posts, Bio, Username)
Data Preprocessing
• Remove missing values
• Normalization
Feature Engineering
• Follower-Following Ratio
• Bio Length
• Username Digit Ratio
Random Forest Classifier (Genuine / Fake)
Account Classification (Genuine / Fake)

Fig. 2: Workflow of Fake Instagram Account Detection

Figure 2 illustrates the workflow used to detect fake Instagram accounts based on profile metadata and Random Forest classification.

3.3 SPAM COMMENT DETECTION MODULE

The spam comment detection module focuses on identifying spam, promotional, or malicious comments. Textual input is preprocessed using standard NLP techniques including lowercasing, removal of special characters, and tokenization. TF-IDF vectorization is applied to convert text into numerical feature vectors. A Logistic Regression classifier is then used to classify the comment as spam or non-spam. This model provides reliable performance while allowing explainable outputs such as important contributing terms.

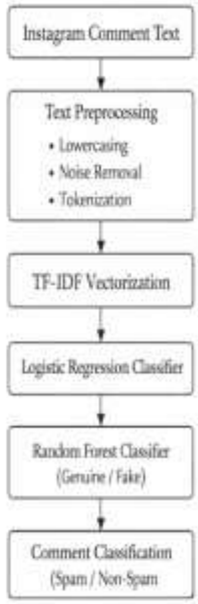
Instagram Comment Text
Text Preprocessing
• Lowercasing
• Noise Removal
• Tokenization
TF-IDF Vectorization
Logistic Regression Classifier
Random Forest Classifer (Genuine / Fake)
Comment Classification (Spam / Non-Spam

Fig. 3: Spam Comment Detection Pipeline

Figure 3 shows the spam comment detection process using text preprocessing, TF-IDF vectorization, and Logistic Regression classification.

4. METHODOLOGY

The methodology adopted in this research follows a structured approach consisting of data collection, preprocessing, feature engineering, model training, and evaluation [4]. The proposed InstaGuard system employs supervised machine learning techniques for both fake account detection and spam comment classification.

Figure 4 illustrates the methodology followed in the proposed InstaGuard system for detecting fake Instagram accounts and spam comments. The process begins with data collection from publicly available Instagram profile metadata and labeled comment datasets. The collected data is then preprocessed to handle missing values, normalize numerical features, and clean textual content. Feature engineering is performed to extract meaningful profile-based attributes and TF-IDF representations for textual data. Subsequently, supervised machine learning models are trained using Random Forest for fake account detection and Logistic Regression for spam comment classification. Finally, the trained models are evaluated using standard performance metrics, and the system generates predictions indicating account authenticity and comment legitimacy.

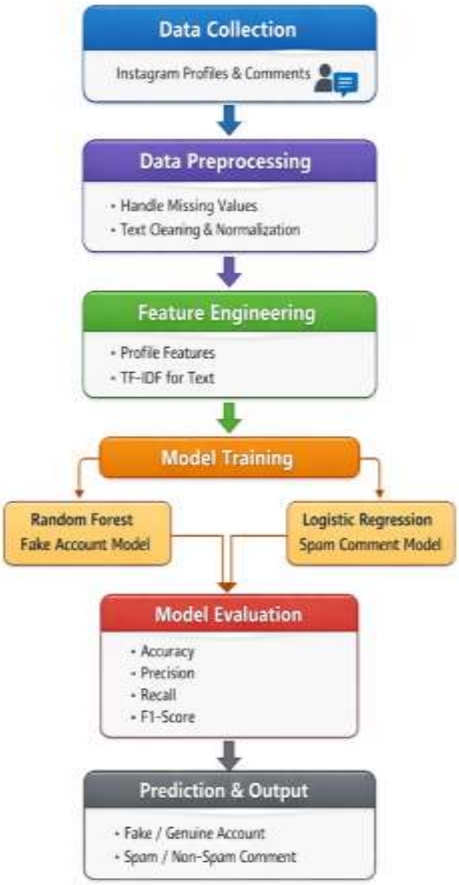

Fig. 4: Methodology of the proposed InstaGuard system illustrating data collection, preprocessing, feature engineering, model training, evaluation, and final prediction of fake accounts and spam comments.

Figure 4 presents the overall methodology followed in the proposed InstaGuard system.

4.1 DATA COLLECTION

The dataset used for fake account detection consists of Instagram profile metadata collected from publicly available sources. The features include follower count, following count, number of posts, profile visibility, biography details, and username patterns. For spam comment detection, a labeled dataset containing spam and non-spam comments was used to train the text classification model.

4.2 DATA PREPROCESSING

For account-level data, missing values were handled and numerical features were normalized where required. Textual attributes such as biography text were cleaned by removing special characters and unnecessary symbols.

For comment-level data, standard natural language processing techniques were applied. These include converting text to lowercase, removing punctuation and URLs, and eliminating extra whitespace. This preprocessing ensures consistency and improves model performance.

4.3 FEATURE ENGINEERING

Feature engineering plays a crucial role in improving classification accuracy. For fake account detection, engineered features include follower-to-following ratio, bio length, presence of external links, and username digit frequency. These features help distinguish genuine users from automated or malicious accounts.

For spam comment detection, TF-IDF (Term Frequency–Inverse Document Frequency) vectorization was applied to convert textual data into numerical feature vectors. This technique captures the importance of words based on their frequency across the dataset.

Feature engineering techniques such as follower-to-following ratio, bio length, username digit frequency, and TF-IDF vectorization are used to improve classification accuracy [3], [5]. Random Forest is employed for fake account detection due to its robustness and ability to handle non-linear relationships, while Logistic Regression is used for spam comment detection because of its interpretability and efficiency [2], [8]. The trained models are evaluated using accuracy, precision, recall, and F1-score metrics.

4.4 MODEL TRAINING

The fake account detection model was trained using a Random Forest classifier due to its ability to handle complex feature interactions and reduce overfitting. The spam comment detection model was trained using Logistic Regression, which provides reliable classification performance and interpretable results.

The datasets were divided into training and testing sets using an appropriate train–test split strategy. Hyperparameters were tuned to optimize classification performance.

4.5 EVALUATION METRICS

The models were evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the effectiveness of the proposed system in detecting fake accounts and spam comments.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation demonstrates that the proposed InstaGuard system effectively identifies both fake Instagram accounts and spam comments using lightweight yet robust machine learning techniques. The system was evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of classification performance.

5.1 FAKE INSTAGRAM ACCOUNT DETECTION RESULTS

The fake account detection module employs a Random Forest classifier trained on profile-based metadata such as follower–following ratio, post count, username characteristics, profile visibility, and biography attributes.

The evaluation results, presented in Table 1, indicate that the model achieves high classification accuracy and reliability.

Table 1: Performance Metrics for Fake Instagram Account Detection

Metric	Value
Accuracy	0.9809
Precision	0.9894
Recall	0.9722
F1-Score	0.9807

The high accuracy (98.09%) reflects the model's strong ability to distinguish between genuine and fake accounts. A precision of 98.94% indicates that the majority of accounts classified as fake are indeed malicious, minimizing false positives. The recall value of 97.22% demonstrates that the model successfully identifies most fake accounts present in the dataset. The high F1-score further confirms a well-balanced trade-off between precision and recall.

These results validate the effectiveness of using ensemble-based models such as Random Forest for handling heterogeneous and non-linear profile features commonly observed in social media account data.

5.2 SPAM COMMENT DETECTION RESULTS

The spam comment detection module utilizes a TF-IDF and Logistic Regression pipeline, evaluated on a balanced subset of the YouTube Spam Collection dataset. The model analyzes textual patterns and term importance to classify comments as spam or non-spam. The evaluation results are summarized in Table 2.

Table 2: Performance Metrics for Spam Comment Detection

Metric	Value
Accuracy	0.926
Precision	0.919
Recall	0.908
F1-Score	0.913

The spam comment detection model achieves an accuracy of 92.6%, indicating reliable overall performance in text classification. The precision value of 91.9% suggests that most comments predicted as spam are correctly identified, while the recall of 90.8% confirms the model's capability to capture the majority of spam comments. The F1-score of 91.3% demonstrates consistent and balanced performance across both evaluation dimensions.

These results show that traditional natural language processing techniques combined with interpretable machine learning models can effectively detect spam content without the computational overhead associated with deep learning approaches.

5.3 COMPARATIVE DISCUSSION

The combined evaluation of both modules highlights the practical strength of the dual-model architecture adopted in InstaGuard. While the fake account detection model achieves slightly higher performance due to structured numerical features, the spam comment detection model also demonstrates strong results despite the inherent complexity of unstructured textual data.

The findings indicate that feature engineering coupled with lightweight machine learning models can deliver high accuracy and interpretability, making the system suitable for real-time deployment. Moreover, the use of confidence scores and transparent predictions enhances user trust and system usability.

Overall, the experimental results confirm that InstaGuard provides a scalable, efficient, and reliable solution for improving safety on social media platforms by simultaneously addressing fake accounts and spam comments.

6. CONCLUSION AND FUTURE WORK

This research presented the design and implementation of InstaGuard, a machine learning-based system for detecting fake Instagram accounts and spam comments. By integrating account-level analysis with comment-level text classification, the proposed system addresses key limitations observed in existing solutions. The use of Random Forest and Logistic Regression models enables effective, interpretable, and scalable detection of malicious activities.

The experimental results demonstrate that the system performs reliably while maintaining simplicity and explainability. The implementation of a real-time web-based dashboard further enhances usability and practical deployment potential.

Future work may include the integration of deep learning models such as transformer-based architectures, multilingual spam detection, emoji and sentiment analysis, and cloud-based deployment for large-scale usage. Incorporating additional contextual features such as post–comment relationships could further improve detection accuracy.

REFERENCES

[1] T. Wu, S. Wen, Y. Xiang, and W. Zhou, “Twitter spam detection: Survey of new approaches and comparative study,” Computers & Security, vol. 76, pp. 265–284, 2018.
Available at:
https://www.sciencedirect.com/science/article/abs/pii/S016740481730250X

[2] A. Chrismanto and Y. Lukito, “Instagram Spam Detection,” IEEE Pacific Rim International Symposium on Dependable Computing, pp. 227–228, 2017.
Available at: https://ieeexplore.ieee.org/document/7920623

[3] S. Aiyar and N. P. Shetty, “N-Gram Assisted YouTube Spam Comment Detection,” Procedia Computer Science, vol. 132, pp. 174–182, 2018.
Available at:
https://www.sciencedirect.com/science/article/pii/S1877050918309153

[4] Duzgun, A., Duran, F., & Ozgur, A. (2018). Comparing Common Supervised Machine Learning Algorithms For Twitter Spam Detection in Scikit Learn. In International Conference on Cyber Security and Computer Science (ICONCS'18) (pp. 18-20).
Available at:
https://indexive.com/uploads/papers/pap_indexive15505850772147483647.pdf

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.
Available at:
https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

[6] A. Vaswani et al., “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), pp. 5999–6009, 2017.
Available at: https://arxiv.org/abs/1706.03762

[7] Rachmat, Antonius, et al. "Spam comments detection on instagram using machine learning and deep learning methods." Lontar Komputer: Jurnal Ilmiah Teknologi Informasi 13.46 (2022): 10-24843.
Available at:
https://www.researchgate.net/publication/362638810_Spam_Comments_Detection_on_Instagram_Using_Machine_Learning_and_Deep_Learning_Methods

[8] Pranali V, Dhote, et al. "Instagram Spam Detection (ISD)." International Journal of Trend in Scientific Research and Development 8.5 (2024): 573-583.
Available at: https://www.ijtsrd.com/other-scientific-research-area/other/69419/instagram-spam-detection-isd/pranali-v-dhote