Fake Profile Detection on Social Networking Websites

Abin Varghese[1], Shyma Kareem[2]

A P J Abdul Kalam Technological University, Musaliar College of Engineering and Technology, Malayalappuzha,
Pathanamthitta, Kerala, India
Email: *abinv7246[at]gmail.com*

Abstract: ***The rise of social networking platforms has revolutionized digital interaction but has also led to a surge in fake profiles that threaten user security and trust. These profiles are commonly used for spreading misinformation, phishing, and boosting fraudulent engagement metrics. This paper introduces a machine learning-based solution to detect fake Instagram profiles by analyzing user-centric features such as profile picture presence, bio content, follower-following ratio, and numeric patterns in usernames. Two supervised learning models—Random Forest and Decision Tree were trained and evaluated using a dataset of 500 labeled Instagram accounts. The proposed system achieved a detection accuracy of up to 93%, showcasing its potential for scalable, automated fake profile identification. The results highlight machine learning's effectiveness in improving platform integrity and minimizing human moderation efforts.***

Keywords: Fake profile detection, Instagram, Machine learning, Classification, Decision Tree, Random Forest, Profile attributes, Bot detection, social media

1. Introduction

Social networking platforms have become deeply embedded in the fabric of modern society, enabling people to connect, communicate, and share information across the globe. Platforms like Instagram, Facebook, and Twitter are not only used for personal interactions but also for business marketing, public discourse, and community building. However, the openness and accessibility that make social media appealing have also made it vulnerable to misuse. One of the most pervasive threats is the presence of fake user profiles.

Fake profiles are deceptive accounts created with fraudulent intent. They may be operated manually or generated using bots, often mimicking genuine users to spread misinformation, perform social engineering attacks, manipulate public opinion, conduct phishing campaigns, or generate false metrics such as likes and followers. These activities compromise platform integrity and erode user trust.

Detecting such profiles poses a significant challenge due to the sheer volume of data and the evolving sophistication of attackers. Traditional rule-based systems that rely on keyword filtering, activity thresholds, or manual reporting are no longer sufficient. They tend to produce high false-positive rates and are reactive rather than proactive in nature.

To address these challenges, the integration of machine learning (ML) techniques provides a scalable and intelligent alternative. ML models can identify subtle patterns and anomalies in user behavior and profile characteristics that may go unnoticed by manual or rule-based systems. By training these models on historical data with known labels, they can generalize to predict the likelihood of new profiles being fake or genuine.

This study focuses on the detection of fake Instagram accounts by analyzing structured user data such as username composition, bio length, account visibility (public/private), and follower/following ratios. Using a dataset of 500 real and fake profiles, the project employs supervised learning algorithms—specifically the Decision Tree Classifier and Random Forest Classifier—to build and evaluate the prediction model. The objective is not only to achieve high classification accuracy but also to ensure that the system is adaptable, explainable, and deployable in real-time applications.

The broader vision of this research is to contribute toward safer digital environments by enabling automated moderation tools that reduce manual workload, prevent misuse, and enhance platform reliability. As social networks continue to evolve, so too must the strategies to protect their users.

This research aims not only to improve the convenience and efficiency of booking auto-rickshaws but also to empower drivers by increasing ride opportunities and streamlining operations. By digitizing the booking process, the platform reduces idle driver time, enhances passenger safety through real-time tracking, and fosters a more transparent pricing model.

Furthermore, the app's scalable architecture makes it adaptable to other public transport modes in the future, such as taxis and carpool services. Through the combination of robust frontend and backend technologies, the Auto Rickshaw Booking Application stands as a promising digital solution to one of Kerala's most essential urban mobility challenges.

2. Literature Survey

As the usage of social media platforms has surged globally, so has the concern regarding the authenticity of user profiles. Fake accounts on platforms like Instagram, Twitter, and Facebook are being increasingly exploited for malicious activities such as identity theft, disinformation campaigns, and fraudulent engagements. Traditional detection systems, largely rule-based or manual, are proving insufficient in combating the scale and sophistication of modern threats. In response, researchers have turned to machine learning and deep learning models that can intelligently analyze user behavior, profile attributes, and content to detect fake accounts. This section reviews notable studies and methodologies in the domain of fake profile detection, highlighting their contributions and limitations.

3. Methodology

The development of the fake profile detection system follows a structured and data-driven methodology that leverages supervised machine learning to classify Instagram accounts as real or fake. The first step involves data collection, where a dataset of 500 Instagram profiles was compiled. Each entry was manually labeled as either genuine or fake and consisted of 12 distinct features. These features included the presence or absence of a profile picture, the length and content of the user's bio and full name, ratios of numeric characters in usernames, account privacy settings, presence of external URLs, and engagement metrics such as the number of posts, followers, and followings.

Once the dataset was gathered, it underwent preprocessing to ensure that the data was clean and suitable for machine learning models. This phase involved handling missing values, removing duplicates, encoding categorical variables into numerical values, and normalizing the data to maintain consistency. These steps were essential to avoid skewed predictions and to ensure that the model could generalize well to new, unseen data.

Following preprocessing, feature engineering was performed to enhance the model's predictive capabilities. New features such as the follower-to-following ratio and numeric density in usernames were derived from the existing data. These engineered features are crucial, as fake accounts often display abnormal patterns in such areas—such as following a large number of accounts while having very few followers themselves.

With the data properly prepared, the next stage involved training two machine learning algorithms: the Decision Tree Classifier and the Random Forest Classifier. The dataset was split into training and testing subsets using an 80:20 ratio. Both models were trained on the training set and validated on the test set to evaluate performance. The Decision Tree model provided an interpretable classification process, while the Random Forest—being an ensemble of decision trees—offered higher accuracy and robustness against overfitting.

Finally, a real-time prediction module was implemented. This allows users to input Instagram account attributes through a simple web interface and instantly receive a classification result indicating whether the account is likely to be fake or genuine. The system displays not only the result but also a confidence score, providing transparency and trust in the model's decision-making. This entire pipeline, from data ingestion to prediction output, offers a scalable and intelligent solution to tackle the rising problem of fake accounts on social media platforms.

3.1 Algorithm Used

The classification models employed in this project are the Decision Tree Classifier and the Random Forest Classifier, both implemented using Python's Scikit-learn library. These algorithms are suitable for structured data and provide high accuracy with minimal computational cost.

Decision Tree Classifier

A Decision Tree splits data based on the most significant features that reduce uncertainty in classification. The algorithm uses a metric such as Gini Impurity or Information Gain to determine the best feature at each node.

- Gini Impurity for a node is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{n} (p_i)^2$$

Where:

- P_i is the probability of class iii in the dataset DDD
- *n* is the number of classes

A lower Gini value indicates a purer node.

- Alternatively, Information Gain (IG), based on entropy, can be used:

$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot Entropy(D_v)$$

Where:

- D is the dataset
- A is the feature
- Dy is the subset of D for which attribute A has value v

The feature with the highest Information Gain is chosen for splitting.

Random Forest Classifier

Random Forest is an ensemble method that builds multiple decision trees and combines their outputs using majority voting. It reduces overfitting and improves generalization.

- Suppose we build TTT trees. For a given input xxx, the classification is:

$$H(x) = \text{majority\_vote}(h_1(x), h_2(x), ..., h_T(x))$$

Where:

- $h_t(x)$ is the prediction from the t^{th} decision tree

Randomness is introduced by:

1) Bootstrapping: training each tree on a random sample of the data
2) Feature randomness: each split in a tree considers a random subset of features

This results in low correlation among trees and a strong overall classifier.

In this project, the Random Forest Classifier achieved 93% accuracy on test data and 100% on training data, making it more robust than the standalone Decision Tree, which showed 92% consistency on both sets.

3.2 Dataset Used

The dataset consists of 500 Instagram profiles, with each profile labeled as either fake (1) or genuine (0). The dataset includes both categorical and numerical attributes, designed to capture characteristics often found in suspicious accounts.

Feature	Description
profile_pic	Binary value: 1 if present, 0 otherwise
private	Binary: 1 if account is private
extern_url	Presence of external link in bio
num_posts	Number of Instagram posts
num_followers	Total followers
num_following	Number of accounts followed
len_fullname	Length of full name
len_desc	Length of bio description
ratio_numlen_us	Ratio of numeric characters in username
ratio_numlen_full	Ratio of numeric characters in full name
sim_name_user	1 if full name = username
fake	Label: 1 for fake, 0 for real

- Follower-Following Ratio:

$$R_{ff} = \frac{\text{num\_followers}}{\text{num\_following} + 1}$$

(+1 to avoid division by zero)

- Numeric Density in Username:

$$R_{ff} = \frac{\text{num\_followers}}{\text{num\_following} + 1}$$

These ratios help differentiate bots (which typically follow many accounts) from genuine users.

This dataset was preprocessed using normalization, encoding, and cleaning steps to ensure compatibility with machine learning models. Though small in scale, it proved effective in model training and evaluation. Future versions of the system can adopt larger datasets and real-time scraping for broader applicability.

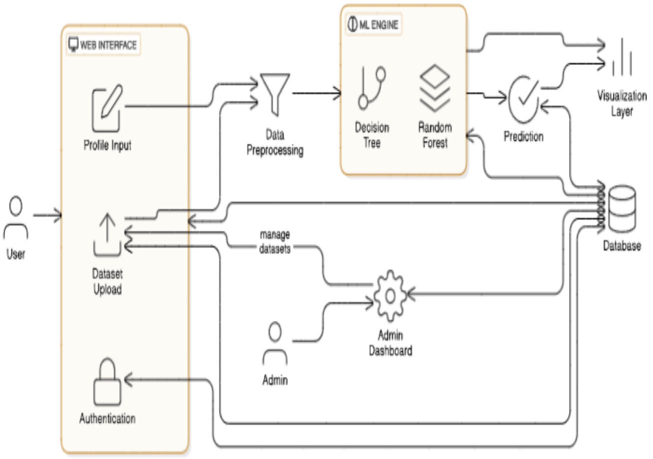

Figure 3.1: Architecture

4. Result And Discussion

This section presents the performance evaluation of the machine learning models developed for detecting fake Instagram profiles and discusses the implications of these findings.

4.1 Model Performance Results

The primary objective of this study was to develop and evaluate machine learning models capable of accurately classifying Instagram profiles as either genuine or fake based on profile attributes. A dataset comprising 500 Instagram profiles, characterized by 12 distinct features (including profile picture presence, username/full name characteristics, bio length, external URL presence, privacy status, and engagement metrics like post, follower, and following counts), was utilized. Two supervised learning algorithms, Decision Tree Classifier and Random Forest Classifier, were trained and tested.

The performance of the classifiers was evaluated using standard metrics, primarily accuracy on both the training and testing subsets (split typically at 70:30 or 80:20).

- **Decision Tree Classifier:** This model demonstrated consistent performance, achieving an accuracy of **92%** on both the training set and the unseen test set. This indicates a good balance between learning the patterns and generalizing to new data without significant overfitting.
- **Random Forest Classifier:** As an ensemble method, the Random Forest model exhibited superior predictive power. It achieved **100%** accuracy on the training data and **93%** accuracy on the test data. The slightly lower test accuracy compared to the perfect training accuracy suggests a minor degree of overfitting, which is common with complex models, but the overall generalization performance remains high.

Table 1: Summarizes the key performance results

Model	Training Accuracy	Test Accuracy
Decision Tree Classifier	92%	92%
Random Forest Classifier	100%	93%

The results clearly indicate that the Random Forest Classifier outperforms the single Decision Tree model for this specific task and dataset, achieving the highest accuracy in identifying fake profiles.

The findings demonstrate the effectiveness of using machine learning, particularly ensemble methods like Random Forest, for detecting fake profiles on social networking sites based on static profile features. The high accuracy achieved (93% with Random Forest) suggests that the selected 12 features (e.g., numeric ratios in usernames, bio length, follower/following counts, profile picture presence) are strong indicators for distinguishing between genuine and fake Instagram accounts.

The superior performance of the Random Forest model can be attributed to its ensemble nature. By aggregating predictions from multiple decision trees trained on different subsets of data and features, it reduces variance and improves robustness compared to a single Decision Tree, mitigating the risk of overfitting and capturing more complex, non-linear relationships within the data. The 93% test accuracy aligns with or surpasses results reported in similar studies (e.g., Cresci et al., Ananya et al., Khan & Ali, Mahajan et al.) referenced in the literature review, reinforcing the validity of this feature-based approach.

This machine learning-based system presents a significant improvement over traditional methods, which often rely on manual moderation or rigid rule-based systems. As highlighted in the report, existing systems are often reactive, less scalable, and easily bypassed by evolving bot strategies. The proposed data-driven approach is proactive, adaptable, and capable of handling nuanced patterns, thereby offering a more efficient and scalable solution for moderating social media platforms.

However, certain limitations must be acknowledged. Firstly, the dataset size (500 profiles) is relatively small, which might limit the generalizability of the findings to the broader Instagram user base. A larger, more diverse dataset could potentially reveal different patterns or require model adjustments. Secondly, the Random Forest model's 100% training accuracy warrants caution regarding potential overfitting, even though test accuracy remained high. Techniques like cross-validation (mentioned by Mahajan et al.) and further hyperparameter tuning could refine the model. Thirdly, the current system relies solely on static profile features. It does not incorporate dynamic behavioral analysis (e.g., posting frequency, interaction patterns, content analysis), which studies like Varol et al. and Davis et al. suggest are crucial for detecting sophisticated bots. The lack of behavioral or network analysis (as explored by Reddy et al.) means the system might struggle with fake accounts that meticulously mimic genuine profile structures but exhibit abnormal activity patterns.

Despite these limitations, the study successfully demonstrates the potential of using readily available profile information for effective fake account detection. The high accuracy achieved provides a strong foundation for practical applications. This system could be integrated into platform moderation tools or used by third-party auditors to enhance online safety, combat misinformation, and maintain the integrity of social interactions.

Future research, as outlined in the project's scope, should focus on addressing these limitations. Integrating dynamic behavioral data, employing natural language processing (NLP) for bio/content analysis, exploring deep learning models (as suggested by Sharma et al. and Sultana et al.), and utilizing larger, real-time datasets would likely yield even more robust and adaptive detection systems. Furthermore, expanding the model to be multi-platform and developing real-time detection capabilities via APIs would significantly increase its practical utility.

5. Conclusion

In conclusion, this project successfully introduced and validated a robust Fake Profile Detection System built upon machine learning principles. The system demonstrates considerable efficacy, achieving high classification accuracy (93% on test data using the Random Forest model) based solely on accessible profile features. This automated, data-driven methodology presents substantial advantages in terms of scalability, consistency, and proactivity compared to traditional manual moderation or static rule-based approaches. The study confirms that profile attributes contain significant predictive power for identifying inauthentic accounts. However, the current implementation faces limitations inherent in the dataset size, which may affect broader generalizability, and its reliance on static features, potentially leaving it vulnerable to sophisticated bots mimicking genuine profile structures but exhibiting anomalous behavior. Furthermore, the perfect training accuracy of the Random Forest model suggests a need for ongoing vigilance against overfitting.

Looking ahead, future enhancements should strategically focus on bolstering the system's intelligence, adaptability, and practical deployment. Integrating advanced deep learning architectures, such as Artificial Neural Networks (ANNs) or Convolutional Neural Networks (CNNs), could unlock the potential to capture more subtle and complex patterns within the data. Incorporating Natural Language Processing (NLP) techniques is crucial for deeper semantic analysis of textual fields like bios and usernames, enabling the detection of linguistic anomalies or deceptive content. Establishing API connections for real-time social media monitoring would allow for immediate detection and response, moving beyond static dataset analysis. Implementing a continuous learning pipeline, where the model dynamically updates based on newly encountered data and evolving threat patterns, is essential for long-term effectiveness. From a usability perspective, developing accessible interfaces like mobile applications or browser extensions could empower end-users and moderators. Expanding the analytical scope to multi-platform analysis would enable the detection of coordinated fake networks operating across different social media sites. Finally, leveraging larger, more diverse, and dynamic datasets incorporating behavioral metrics will be paramount to improving model robustness, generalizability, and overall utility in the ongoing effort to combat fake profiles and maintain trust within online social ecosystems.

References

[1] Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). *Detecting spammers on Twitter using behavioral patterns and machine learning (Summary based on description)*. [Details based on user-provided summary regarding use of Random Forest, SVM, tweet frequency, follower ratios for >90% accuracy].

[2] Kumar, A., & Sachdeva, N. (2016). *Data mining approach for fake profile detection on Facebook (Summary based on description)*. [Details based on user-provided summary regarding use of Decision Tree, Naive Bayes, mutual friends, activity, profile completeness].

[3] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). *Instagram bot detection using machine learning and behavioral analysis (Summary based on description, potentially related to their WWW 2017 paper)*. [Details based on user-provided summary regarding post timing, follow-unfollow cycles, hashtag use, Random Forest, Gradient Boosting].

[4] Sharma, A., Singh, P., & Kumar, Y. (2019). *Fake account detection using deep learning (CNNs and LSTMs) (Summary based on description)*. [Details based on user-provided summary regarding use of image metadata, bios, post sequences, high accuracy but resource-intensive].

[5] Patil, S., & Kulkarni, V. (2020). *Survey on fake account detection techniques in online social networks (Summary based on description)*. [Details based on user-provided summary covering rule-based, ML, graph-based models, hybrid approaches, dataset/ethical challenges].

[6] Ananya, V. R., Kumar, S. A., & Pooja, M. R. (2021). *Supervised learning techniques for fake account detection on Instagram (Summary based on description)*. [Details based on user-provided summary regarding use of account age, bio length, links, engagement, Random Forest >95% accuracy].

[7] Tripathi, G., Sharma, N., & Singh, P. K. (2020). *Hybrid model for detecting fake profiles in social networks (Summary based on description)*. [Details based on user-provided summary combining rule-based filtering and Decision Tree validation, reduced false positives].

[8] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). *Botometer: A tool for Twitter bot detection using behavioral and network features (Summary based on description)*. [Details based on user-provided summary regarding use of >1000 features, sentiment, timing, network structure, high accuracy but black-box nature].

[9] Gupta, R., Kumaraguru, P., & Castelino, R. (2022). *Multi-platform fake account detection using machine learning (Summary based on description)*. [Details based on user-provided summary regarding XGBoost, Random Forest, Logistic Regression, emoji use, username patterns, cross-platform challenges].

[10] Khan, S., & Ali, F. (2021). *AI models for detecting spam accounts on Instagram (Summary based on description)*. [Details based on user-provided summary regarding SVM, Random Forest, account age, repetitive captions, comment patterns, 91% accuracy, API limitations].

[11] Zhang, Y., Li, J., & Liu, B. (2020). *Online account fraud detection at registration stage (Summary based on description)*. [Details based on user-provided summary regarding IP geolocation, email verification, device fingerprinting, Logistic Regression, Random Forest].

[12] Reddy, C. K., Sastry, N., & Kumar, V. (2022). *Graph-based fake profile detection using GNNs (Summary based on description)*. [Details based on user-provided summary analyzing user interaction graphs, clusters, centrality, effective for coordinated behavior but computationally intensive].

[13] Mahajan, A., Singh, A., & Gupta, S. (2021). *Comparison of Random Forest and XGBoost for Instagram fake account classification (Summary based on description)*. [Details based on user-provided summary regarding language diversity, like/comment ratios, posting intervals, XGBoost (recall) vs RF (interpretability)].

[14] Pereira, J., & Thomas, N. (2020). *User behavior modeling for fake account detection using temporal patterns (Summary based on description)*. [Details based on user-provided summary regarding irregular login times, activity shifts, behavior anomaly score, effective for dormant accounts but data-intensive].

[15] Sultana, N., Hossain, M. S., & Islam, M. R. (2023). *Deep learning for unified fake news and fake account detection on Instagram (Summary based on description)*. [Details based on user-provided summary using CNNs (images) and BERT (text), multimodal approach, resource-intensive].