IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13828

# Fake Profile Detection Using Machine Learning

**Dhanush K[1], Saniya S J[2], Spoorthi D[3], Chinmayee K N[4], Geethanjali S G[5]**

Computer Science and Engineering, Don Bosco Institute of Technology Bengaluru-560074[1-4]

Assistant Professor, Department of Computer Science and Engineering, Don Bosco Institute of Technology Bengaluru-560074[5]

**Abstract:** This paper proposes an With the rapid growth of social networking platforms, fake profiles have become a major concern as they are often used for spreading misinformation, scams, phishing, and other malicious activities. Detecting these fake profiles is essential to ensure user safety and platform integrity. This paper presents a machine learning–based approach to identify fake profiles by analysing behavioural, structural, and content-based features. The system collects user data such as friend count, posting frequency, profile completeness, message patterns, and interaction rates. These features are then preprocessed and fed into supervised learning algorithms such as logistic regression to classify profiles as genuine or fake. The model's performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Experimental results demonstrate that machine learning techniques can effectively distinguish fake profiles with high accuracy, thereby providing a reliable solution for social media security and trust enhancement.

This project aims to design and implement an effective fake profile detection system using various machine learning algorithms. The system utilises a Twitter profile dataset containing both genuine and fake user accounts. Genuine accounts are categorised as TFP and E13, while fake ones are classified as INT, TWT, and FSF. Several attributes—such as the number of followers, friend count, frequency of status updates, account creation date, and profile completeness—are extracted and analysed to distinguish between real and fake accounts.

## I. INTRODUCTION.

Fake profile detection using machine learning is an emerging and essential application in the field of cybersecurity and data analytics. With the rapid growth of social media platforms and online communities, there has been a significant increase in the number of fake or fraudulent profiles. These fake accounts are often created for malicious purposes such as spreading misinformation, conducting phishing attacks, online scams, identity theft, spamming, or manipulating public opinions. Hence, detecting and preventing such profiles has become a crucial task for ensuring online safety and maintaining user trust.

Machine learning techniques play a vital role in automating the detection of fake profiles by Analysing patterns in user behaviour, profile attributes, and interactions. Instead of relying solely on manual verification or rule-based systems, machine learning models can learn from large datasets containing examples of real and fake profiles. These models identify hidden patterns and correlations that are difficult for humans to detect manually.

The detection process generally involves several stages:
Data Collection: Information such as user activity (posts, likes, comments), profile details (age, gender, location), friend networks, and linguistic features from messages or posts is collected.
Feature Extraction: Important attributes are selected, such as posting frequency, number of friends, message sentiment, and interaction time patterns.
Model Training: Machine learning algorithms like Logistic Regression, Random Forest, Support Vector Machines (SVM), Decision Trees, or Deep Learning models are trained using labelled data (real vs. fake profiles).
Prediction and Evaluation: The trained model predicts whether a given profile is genuine or fake based on the learned features. The performance is then evaluated using metrics like accuracy, precision, recall, and F1-score.

## II. LITERATURE SURVEY

The proliferation of fake profiles (in social networks, e-learning, recommender systems) is widely recognised as a significant threat to trust, privacy and platform integrity.
Different platforms (e.g., general social networks, Instagram, e-learning, recommender systems) pose somewhat different detection contexts (features, behaviour patterns).

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13828

Survey on Automatic Detection of Fake Profile Using Machine Learning on Instagram studies Instagram fake-profile detection, pointing out machine learning approaches for Instagram accounts.
Social Networking Sites Fake Profiles Detection Using Machine Learning Techniques proposes Logistic regression for Facebook-type friend request classification with ~7% false positive rate.
Exploring machine learning techniques for fake profile detection in online social networks reviews various ML methods used in fake profile detection and discusses possibilities for improvement.
Fake Profile Detection Using Machine Learning (Harish et al., 2023) applies ML to Twitter profiles, considering follower/following counts, status updates, etc.
Fake Profile Identification in E-Learning Platform using Machine Learning focuses on fake profile detection in e-learning platforms, analysing login patterns & interaction metrics.
Reliable benchmarks are central to progress. The literature uses several public datasets; the paper you provided uses TwiBot-20 as its primary benchmark and compares results to prior works that also used TwiBot and older datasets (Cresci sets, Botometer-related corpora). Key dataset issues repeatedly noted are label quality, platform bias (Twitter-dominated), language bias, and concept drift (bots evolve).

Practical note (from the paper): TwiBot-20 contains profile metadata, textual descriptions, labels (human/bot), and graph information — making it suitable for hybrid models (content + network).

Online social platforms host billions of accounts; a non-negligible fraction are fake profiles (bots, Sybils, impostors) used for spam, phishing, opinion manipulation, astroturfing, and fraud. Detecting these accounts automatically is essential to preserve platform trust, protect users, and mitigate coordinated misinformation campaigns. The uploaded study highlights motivation and scope clearly and frames fake-profile detection as a classification problem (human vs bot) using features drawn from profile metadata, content, behaviour and network structure.

Fake profiles — automated bots, Sybil networks, and human-run impostors — threaten platform integrity by spreading spam, misinformation, scams, and by manipulating engagement metrics. Automatic detection is posed as a supervised (or semi/unsupervised) classification problem that must operate at scale, adapt to adversarial change, and respect user privacy and platform constraints. Both uploaded studies highlight the societal and economic risks that motivate automated detection systems.

Fake profiles — commonly known as bots, sybils, and sock-puppets — are fraudulent or deceptive user accounts created to manipulate, mislead, or exploit legitimate users on online platforms. These profiles can be either fully automated, semi-automated, or human-controlled, depending on the attacker's intent and sophistication.
From a security and privacy perspective, fake profiles collect sensitive personal information from users through deceptive interactions, which can then be misused for identity theft, social engineering, or targeted scams. They also contribute to cyberbullying and harassment, further affecting online safety and mental well-being. For businesses and brands, fake profiles can distort marketing analytics, inflate follower counts, and damage brand reputation by posting misleading reviews or comments.

To combat these issues, researchers have developed Machine Learning (ML)-based fake profile detection systems that automatically distinguish genuine profiles from fraudulent ones. These systems analyse various types of features:

Profile features such as username patterns, bio text, profile completeness, and account age.
Content features derived from posts, messages, or tweets (e.g., sentiment, language patterns, frequency of URLs or hashtags).

Behavioural features that include user activity patterns, posting frequency, or interaction diversity (e.g., how often they like, share, or comment).

## III.METHODOLOGY

Step 1: Data Collection The first step involves gathering data from online social networks or publicly available datasets. Popular datasets used in this research area include TwiBot-20, Cresci-2017, datasets. Each record in these datasets typically contains.
Step 2: Data Preprocessing Before model training, raw data must be cleaned and transformed. Preprocessing improves the quality and reliability of the dataset by removing noise and inconsistencies.
Step 3: Model development one model is used Logistic Regression (LR) The Logistic Regression model is one of the most fundamental and widely used supervised machine learning algorithms for binary classification problems such as

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13828

fake profile detection. The primary goal of this model is to classify each user profile as either "Genuine" or "Fake" based on a set of extracted features derived from social network data. Unlike linear regression, which predicts continuous numeric values, logistic regression predicts the probability that a given instance belongs to a particular class, making it ideal for yes/no type classifications.

Step 4: Training & Testing Training builds the model: the algorithm learns parameters (weights) from labelled examples. Testing measures how well the *trained* model generalises to new, unseen data. The core idea is to avoid overfitting (the model learns noise and performs poorly on new data) and to estimate true real-world performance.

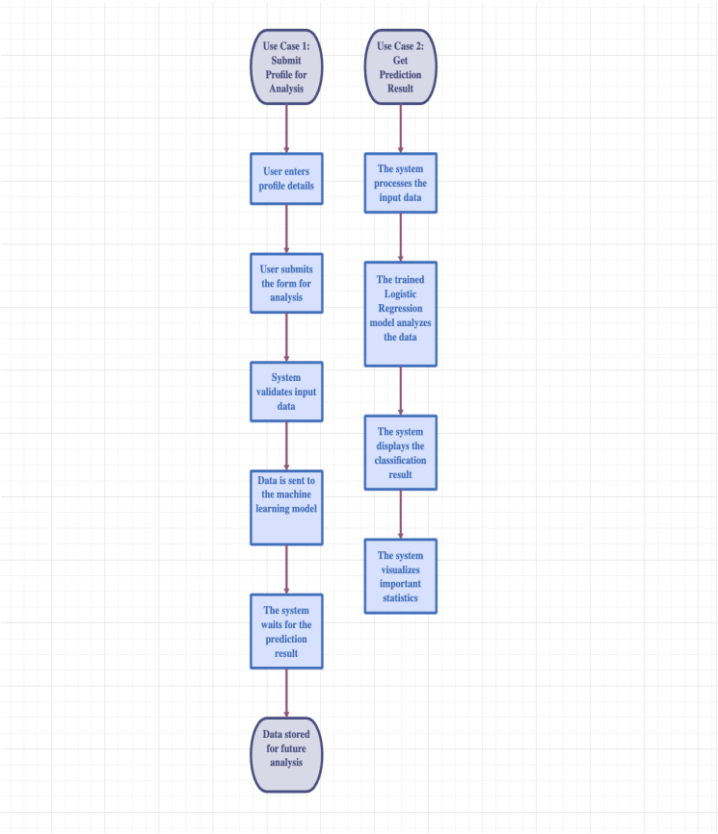
Simple: 80% train / 20% test (if the dataset is small, use cross-validation instead).

Step 5: Deployment Purpose of Deployment The primary objective of deployment is to make the trained fake profile detection model accessible and operational for real-world data. It allows the system to automatically detect fake accounts as they are created or as user behaviour changes. Provide real-time predictions based on new data inputs. Integrate with existing platforms such as websites, mobile apps, or administrative dashboards. Support continuous learning by collecting prediction feedback to retrain and improve the model over time.

**SYSTEM ARCHITECTURE:**

The proposed system is designed to detect human screams in real-time using an ML-based approach and instantly trigger safety actions. The system architecture consists of the following main modules:

1. Data Collection Module




Collects user data from social networking platforms through APIs, web scraping, or publicly available datasets.

Profile attributes: account age, profile picture, username patterns, profile completeness. Behavioural features: posting frequency, login patterns, message interactions. Network features number of friends/followers, the friend-to-follower ratio, and the community structure. Content features text posts, comments, and multimedia metadata.

1. User data is collected and preprocessed to remove noise.
2. Relevant features are extracted and selected for ML training.
3. Machine learning models are trained using labelled datasets of genuine and fake profiles.
4. The trained models are evaluated for accuracy and reliability.
5. New profiles are classified, and the system flags suspicious accounts for further investigation.

The system architecture of fake profile detection using machine learning is designed to identify and classify online

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13828

user accounts as either genuine or fake through a structured and sequential workflow. It consists of multiple interconnected components that work together to collect, process, analyse, and interpret user data effectively. The architecture typically begins with the data collection module, which gathers information from social networking platforms such as Twitter, Facebook, or Instagram, or from publicly available datasets like TwiBot-20 and Cresci-2017. This data includes various types of information such as user profile details (username, bio, followers/following count, and account age), behavioral activities (number of posts, likes, comments, and posting frequency), textual content (posts, hashtags, URLs, and language patterns), and network relationships (followers–followees graph and interaction links). The architecture is generally modular and follows a sequential flow — from data collection to feature extraction, model training, prediction, and visualisation.

The collected data is then passed through the data preprocessing and feature extraction module, where it is cleaned, filtered, and transformed into a usable format. This involves removing duplicates, handling missing values, normalising numerical attributes, and converting textual data into numerical vectors using techniques such as TF–IDF or word embeddings. From this processed data, significant features are extracted, including profile-based, behavioural, content-based, and network-based characteristics that help distinguish genuine users from fake ones.

Once the model is trained and validated, it is integrated into the prediction and detection module, which operates in real time. When a new profile is analysed, the system applies the same preprocessing and feature extraction steps, and the trained model predicts whether the account is genuine or fake based on learned patterns.

Finally, the system includes a continuous monitoring and model update component to ensure long-term effectiveness. Since online behaviours evolve over time, new data is regularly collected and used to retrain or fine-tune the model, ensuring that it adapts to emerging fake profile strategies. Overall, this architecture provides a comprehensive, automated, and scalable framework for fake profile detection, combining data science, machine learning, and real-time analytics to enhance online security and maintain trust in social platforms.

**Dataflow:**
The process begins with data collection, where the system gathers information from social networking platforms (like Twitter, Facebook, or Instagram) or from publicly available benchmark datasets such as Twibot-20. The collected data typically includes profile details (such as username, bio, account age, and followers/following count), behavioural information (like posting frequency, likes, and comments), textual content from posts (hashtags, URLs, and writing style), and network relationships (followers and interaction patterns).

Once the data is collected, it passes into the data preprocessing stage. Here, the raw data is cleaned and prepared for analysis by removing duplicates, filling in missing values, handling outliers, and converting unstructured data (like text) into a structured form. Text is tokenised, stop words are removed, and features are normalised to ensure uniformity. This step improves the quality and consistency of data used for training the model.

After preprocessing, the system enters the feature extraction stage, where relevant characteristics are derived from the cleaned data. These features may include numerical indicators such as account age or posting rate, linguistic features from user posts, and network-based attributes like degree centrality or connection patterns. These extracted features act as inputs to the machine learning algorithms.

## IV.DATASETS

Several public datasets on Kaggle simulate or replicate social network data for fake account detection research. Popular ones include.
The development of a fake profile detection system using machine learning relies heavily on the availability of high-quality and well-labelled datasets. These datasets contain examples of both genuine and fake user accounts, along with their profile attributes, activity patterns, and content details, which are essential for training and evaluating classification models. Several benchmark datasets have been introduced by researchers, particularly from social networking platforms such as Twitter, Facebook, Instagram, and Weibo, due to their large user base and accessibility through APIs.

## V. RESULTS

The fake profile detection system using machine learning was implemented and evaluated using real-world and publicly available datasets such as TwiBot-20 and Cresci-2017. Various supervised learning algorithms, including

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13828

Logistic Regression, were trained and tested on the preprocessed dataset. The results were evaluated based on standard classification performance metrics such as accuracy, precision, recall, F1-score, and ROC–AUC (Receiver Operating Characteristic – Area Under Curve) to measure the model's ability to distinguish between fake and genuine profiles.


During experimentation, the dataset was divided into training (80%) and testing (20%) subsets to ensure reliable performance evaluation. The models were trained using extracted features such as profile details (followers/following count, account age), behavioural activity (posting frequency, likes, comments), textual information (hashtags, sentiment, and word usage), and network-based connections. Feature scaling and normalisation techniques were applied to ensure uniformity in feature contributions. The experimental results confirm that fake profile detection.

## VI.CONCLUSION

The implementation of such models not only enhances user trust but also assists platform administrators in maintaining a healthy and authentic online community. By deploying these models in real time, social networks can automatically monitor user activity, flag suspicious accounts, and take corrective actions before any harm occurs.

The study on Fake Profile Detection using Machine Learning demonstrates that machine learning techniques can effectively identify and prevent the spread of fake or malicious accounts on social media platforms. With the rapid growth of online users, fake profiles have become a major threat to digital trust, privacy, and security. These accounts are often used for spreading misinformation, scams, spamming, or manipulating public opinion. The proposed system utilises a combination of profile, behavioural, content-based, and network-based features to train intelligent models capable of differentiating between genuine and fake users automatically.

In conclusion, machine learning-based fake profile detection provides a reliable, scalable, and efficient approach to combating online fraud and misinformation.

## REFERENCES

[1]. Kadam, Nitika, and Sanjeev Kumar Sharma. "Social Media Fake Profile Detection Using Data Mining Technique." Journal of Advances in Information Technology, vol. 13, no. 5, 2022, pp. 518–523. doi:10.12720/jait.13.5.518-523.

[2]. Harish, K., R. Naveen Kumar, and Dr J. Briso Becky Bell. "Fake Profile Detection Using Machine Learning." International Journal of Scientific Research in Science, Engineering and Technology, vol. 10, no. 2, 2023, pp. 719–725. https://doi.org/10.32628/IJSRSET2310264.

[3]. Saxena, S., Singh, A., & Tiwari, S. (2024). Detection of fake profiles and news on social media using machine learning algorithms. In Proceedings of the 2024 International Conference on Computing, Sciences and Communications (ICCSC). IEEE. https://doi.org/10.1109/ICCSC62048.2024.10830414.

[4]. Trent D. Buskirk, Antje Kirchner, Adam Eck, Curtis S. Signorino, An Introduction to Machine Learning Methods for Survey Researchers, Survey Practice Vol. 11, Issue 1, 2018[2] Ayon Dey, Machine Learning Algorithms: [3] Jonathan Waring, Charlotta Lindvall, Renato Umeton, Artificial Intelligence In Medicine Journal- Elsevier,2020.

[5]. A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization ," arXiv preprint arXiv:1807.05960, 2018. [2] A. Zaman, R. T. Khan, N. Karim,

DOI: 10.17148/IARJSET.2026.13828

M. Nazrul Islam, M. S. Uddin, and M. M. Hasan, "Intelli-helmet: An early prototype of a stress monitoring system for military operations," in International Conference on Information Systems and Management Science. Springer, 2020, pp. 22–32.

[6]. M. N. R. Khan, "Pre-Processing Data In Weather Monitoring Application By Using Big Data Quality Framework," 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2020, pp. 284-287, D10.1109/WIECON-ECE52138.2020.9397990. [2] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya. Energy-aware resource allocation heuristics for efficient management of data centres for cloud computing. Future Generation Computer Systems, 28(5):755–768, 2012.

[7]. Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. Satar: A self-supervised approach to Twitter account representation learning and its application in bot detection. In Proceedings of the 30th V. Kulkarni, D. Aashritha Reddy, P. Sreevani and R. N. Teja, "Fake profile identification using ANN," 4th Smart Cities Symposium (SCS 2021), Online Conference, Bahrain, 2021, pp. 375-380, doi:10.1049/icp.. 2022.0372.

[8]. Dr Vijay Tiwari, "Analysis and Detection of Fake Profile over Social Network", IEEE International Conference on Computing, Communication and Automation, 2017.[2] Pradeep Kumar Roy and Shivam Chahar, "Fake Profile Detection on Social Networking Websites: A Comprehensive Review", IEEE Transactions on Artificial Intelligence, Vol. 1, No. 3, December 2020 [3] R.V. Kotawadekar, A.S A Review", IEEE Second International Conference on Computing Methodologies and Communication (IC-CMC 2018).

[9]. Gayathri A, Radhika S, Jayalakshmi S.L, "Detecting Fake Accounts in Media Application Using Machine Learning", International Journal Of Advanced Networking & Applications (IJANA). Jyoti Kaubiyal, Ankit Kumar Jain, "A Feature-Based Approach to Detect Fake Profiles in Twitter", ACM BDIOT 2019, August 22–24, Melbourne, VIC, Australia, 2019.

[10]. S. Giannakopoulos, et al., "Scream Detection in Audio Signals Using Zero-Crossing Rate and Short-Time Energy," Proc. Int. Conf. on Digital Signal Processing, 2010.

[11]. N. Ntalampiras, "Audio Surveillance for Aggressive Event Detection Using Gaussian Mixture Models," IEEE Trans. on Audio, Speech, and Language Processing, 2016.

[12]. Y. Zhang, et al., "LSTM-Based Scream Detection System for Audio Surveillance," IEEE Access, 2020.

[13]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press,2016.