Cross-Platform Fake Profile Detection in Social Media for Misinformation and Cybercrime Prevention

Sumit B.Dhande*
*Research Scholar P.G Department of Computer Science and Engineering SGBAU Amravati, India

Dr. Swati S. Sherekar*
P.G Department of Computer Science and Engineering SGBAU, Amravati, India

Abstract
Digital social networks are now becoming an essential part of interpersonal connections and interactions, information exchange, digital identity formation. On the other hand, their rapid expansion has also enabled the widespread creation of fake profiles used for misinformation campaigns, financial fraud, phishing attacks, identity theft, and coordinated influence operations. These accounts are often operated either by automated bots or organized human networks and increasingly function across multiple platforms simultaneously. Most existing detection systems are platform-specific and fail to capture cross-platform behavioural patterns exploited by modern threat actors.
This article outlines an extensive review of fraudulent accounts detection across multiple online platforms to overcome hurdles regarding misinformation and cybercrime. Various approaches, including machine learning, deep learning, graph-based, and hybrid strategies are analyzed and compared based on performance, scalability, and adaptability.The study highlights key limitations such as lack of cross-platform generalization, dataset constraints, and limited interpretability. A unified framework is discussed to improve detection efficiency and robustness. The findings provide insights into current trends and future research directions for secure and scalable social media analysis.

Keywords: Fake Profile Detection, Cross-Platform Systems, Cybercrime Prevention, Social Network Security, Deep Learning, Misinformation Detection.

I.Introduction

Social networking sites including Instagram, Twitter (X), TikTok, and Reddit have substantially redefined global communication and knowledge acquisition [1], [2]. While these platforms provide social, political, and economic benefits, they also present opportunities for misuse. Dummy profiles and automated accounts are now widely employed to propagate disinformation, manipulate public opinion, conduct frauds, and illegal surveillance [3,5].As digital interactions increasingly influence real-world outcomes, detecting malicious or deceptive accounts has become a critical concern in cybersecurity and social computing research [6, 7]. Initially identification initiatives counted mostly on rule-based filters and hands-on screening. However, the volume and sophistication of fake accounts have rendered these approaches insufficient [8].To address this challenge, authors have utilized machine learning and deep learning techniques that analyze behavioural patterns, linguistic signals, metadata attributes, and network structures [9,12]. Graph-based learning models and hybrid architectures combining multiple feature modalities have demonstrated improved detection accuracy. Despite these advancements, practical deployment remains challenging due to issues such as scalability, adversarial adaptation, privacy limitations, and limited model interpretability[7,12].By leveraging recent developments in deep learning and hybrid modelling techniques, the proposed work aims to contribute toward enhancing the reliability and security of online social platforms [13,14].This review paper showcases a comparative study of prior forged

DOI: https://doi.org/10.5281/zenodo.20832761

profile recognition approaches used in online communities. The study examines various detection techniques and evaluates their effectiveness for cross-platform deployment, misinformation mitigation, and cybercrime prevention based on findings reported in recent literature.

This paper stems from the surging extensiveness of imposter and automated profiles across major social networking websites and their adverse impact on digital trust and platform integrity. Recent surveys and empirical studies highlight that existing detection mechanisms struggle to cope with evolving adversarial behaviours, large-scale data volumes, and platform-specific constraints [6, 7, 9]. Additionally, the growing use of sophisticated automation tools and coordinated bot networks has further exposed the limitations of traditional detection approaches, motivating the need for intelligent and scalable solutions [8,11].Fake profile detection in multimedia-rich digital websites presents additional obstacles because of the diversity and wide range of data involved [17]. These challenges collectively inspired the formulation of this paper.

The paper focuses on the rapid and widespread adoption of social media platforms for personal communication, business promotion, and information dissemination. While these platforms have enhanced global connectivity, they have simultaneously witnessed a significant rise in fake profiles, automated bots, and malicious accounts. Existing literature highlights that such fake entities are increasingly used for misinformation propagation, online fraud, phishing, and manipulation of public opinion, thereby threatening the trustworthiness of digital ecosystems [29,36]. Early studies relied on basic behavioral and content-based analysis, but the growing scale and sophistication of fake profiles have rendered traditional approaches inadequate. This gap between emerging threats and existing detection capabilities has motivated the study of an intelligent, scalable, and learning-based fake profile detection framework.

II.Literature Review

This section shows a summary of prior studies done by researchers related to imposter profile identification across virtual communities. It examines the methodologies, datasets, performance outcomes, and limitations of existing approaches proposed by different researchers. This review helps to pinpoint the ongoing investigations patterns, challenges, and potential directions for future improvements in cross-platform fake profile detection systems.

- **Computer vision techniques, including Scale Invariant Feature Transform (SIFT).**
  Early research on Instagram remains limited when compared to other social media platforms, despite its rapid expansion and influence. One of the foundational studies by Hu et al. (2014) [2] addressed this gap by introducing a methodological framework that combined qualitative and quantitative analysis to examine Instagram content and user behaviour. The study employed computer vision techniques, which include Scale Invariant Feature Transform (SIFT), to extract and organize visual characteristics from large collections of user-posted images.

- **Extreme Gradient Boosting (XGBoost)**
  B. Ghosh et al.(2024)[3] has done in-depth study highlighting a clear methodological transition toward advanced machine learning and deep learning techniques for finding deceptive activities on social websites. The study observes that earlier research predominantly relied on classical classifiers such as Support Vector Machines (SVM), often combined with handcrafted or hybrid feature representations. In contrast, recent approaches increasingly leverage deep learning architectures, including Convolutional Neural Networks (CNNs), along with ensemble learning approaches like Extreme Gradient Boosting (XGBoost), to achieve improved classification performance.

- **Machine Learning and Image-Based Fake Account Detection**
  Ezarfelix et al. (2022) [7] reviewed commonly adopted methodologies for discovering bogus accounts on digital hubs, highlighting the widespread use of supervised machine learning algorithms such as Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines. Their findings are focused on the importance of user metadata in training predictive frameworks for identifying fake accounts. In addition to metadata-based classification, the authors discussed the growing use of computer vision techniques, particularly neural networks, to analyze visual content

DOI: https://doi.org/10.5281/zenodo.20832761

associated with social media profiles, including profile images, posted media, liked images, and tagged content. The review demonstrated that integrating metadata-driven machine learning with image-based deep learning approaches enhances the accuracy of fraud profiles spotting.

- **Hybrid Ensemble Learning for Fake Profile Detection**

  Chakraborty et al. (2022) [23] investigated fake profile and spam account uncovering using a variety of machine learning algorithms, including neural networks, SVM, KNN, and decision trees, based on user profile and behavioural features. The study addressed common challenges such as class imbalance and adaptive fake account strategies by adopting a hybrid ensemble methodology that integrated XGBoost for feature learning and Random Forest for classification robustness. By employing SMOTE for data balancing and GridSearchCV for hyper parameter optimization, the proposed framework achieved improved detection accuracy and generalization. This work highlighted the effectiveness of ensemble-based and optimized learning approaches for scalable and socially responsible fake profile detection.

- **Neural Network–Based Instagram Fake Profile Detection**

  Harish et al. (2023) [26] examined the use of machine learning techniques to recognize scam profiles on Instagram, addressing the growing misuse of the platform for scams and fraudulent promotions. By training neural network models to analyze user profile attributes, the researcher successfully separated authentic accounts from fraudulent ones. The system achieved accuracy rate over 90% during testing, underscoring the capability of neural networks in automating the identification of dummy profiles. This study reinforced the significance of intelligent learning models in securing rapidly growing social media platforms.

- **Majority Voting–Based Fake Social Media Profile Detection**

  Patil et al. (2024)[41] To tackle the issue of fake social media accounts, the researcher built an ensemble learning framework centred around a majority voting system. Instead of relying on just one algorithm, they blended a diverse mix of classifiers including Decision Tree, Random Forest, XGBoost, AdaBoost, Logistic Regression, Extra Trees, and K-Nearest Neighbours to make their predictions much more robust. When tested against the MIB dataset, this combined approach delivered exceptional results, hitting a flat 99.12% across accuracy, precision, recall, and F1-score, outperforming several individual classifiers. This work demonstrated that ensemble methods can provide scalable and highly effective solutions for securing online social platforms.

- **Deep Learning–Based Heterogeneous Social Media Fake Profile Detection**

  Aditya and Mohanty (2024)[39] presented a deep learning framework for tracking down counterfeit accounts across heterogeneous social networking websites using multimodal data sources. The model integrated textual, visual, behavioural, and profile-based features with optimization techniques to enhance classification performance. When merged datasets were used for testing the model, it showed high detection accuracy and better productivity than existing ones. The findings underscore just how crucial multimodal deep learning has become for countering fake profile threats across various online platforms.

- **AI-Based Multi-Model Fake Profile Detection for Platform Credibility**

  Chakranarayan et al. (2025)[38] proposed an AI-driven multi-model framework to detect fake profiles and protect brand as well as platform credibility on social media networks. The framework relies on a three-pronged approach, it uses RoBERTa for textual insights, ConvNeXt for image processing, and Heterogeneous Graph Attention Networks to model how users interact with one another. Experimental results on benchmark datasets achieved 98.9% accuracy, demonstrating strong capability against deceptive and coordinated fake accounts. This research emphasized the value of multimodal artificial intelligence in strengthening trust, security, and reputation management across digital platforms.

- **Multi-Model Joint Representation for Fake User Detection**

  Li et al. (2024)[40] introduced a multi-model joint representation framework for detecting fake users by combining multiple feature modalities into a unified learning model. The

DOI: https://doi.org/10.5281/zenodo.20832761

study integrated textual information, user behaviour, and profile characteristics to capture complex patterns associated with fraudulent accounts. Experimental results showed a significant improvement in detection accuracy compared with conventional single-model approaches. This research demonstrated that joint representation learning can effectively enhance fake user identification in dynamic interactive media environments.

- **Hybrid Deep Learning and Optimization for Fraudulent Account Detection**
  To improve fraudulent account detection on social media, Shukla et al. (2024) designed a multi-layered framework that combines a Temporal Convolutional Network with GAN-driven data augmentation and Autoencoder feature reduction. To maximize the system's learning capacity, the authors integrated the Seagull Optimization Algorithm for precise hyperparameter tuning. The approach was validated using the Cresci-2017 and TwiBot-22 benchmarks, yielding high precision and F1-scores alongside strong ROC-AUC results of 0.96 and 0.95, respectively. Ultimately, the research highlights how blending sequence modelling with synthetic data generation and metaheuristic optimization can create a highly scalable defence against fraud accounts.

III.Existing Methodology

A. Data Sources

The study will rely on publicly available and benchmark social media datasets such as FakeNewsNet, Twitter Bot,Weibo and Twitter, Facebook, Instagram datasets related to fake profile and bot detection. These include datasets containing user metadata, behavioural attributes, textual content, and network connections from platforms such as Twitter and Instagram.

In addition, cross-platform data collected through ethical web scraping and APIs will be used to analyze coordinated fake account behavior.Where necessary, custom datasets will be prepared through controlled data collection and labelling processes to address class imbalance and platform-specific variations, ensuring realistic and representative evaluation scenarios.

Among these, behavioural and network-based signals are particularly valuable in cross-platform environments because they capture structural patterns that remain relatively consistent across different platforms [6, 20].

B. Machine Learning Methods

Traditional machine learning still plays a massive role in spotting fake accounts, mostly because it excels at handling structured data like follower counts, posting habits, and engagement metrics. Algorithms such as Random Forest, SVM, Decision Trees, and Logistic Regression are the go-to choices here, offering a great balance of speed and clear, interpretable logic. While simpler than deep learning, these models work incredibly well on standard tabular datasets and provide the essential baseline needed to prove whether a newer method is actually better. These machine learning techniques have also been applied effectively for detecting fake Instagram users using supervised classification methods [16].Machine learning frameworks tailored for digital platforms like Instagram have shown promising results in identifying fraudulent accounts [21].

C. Deep Learning Approaches

Deep learning approaches are adopted to learn hidden patterns from large-scale and unstructured social media data. Convolutional Neural Networks are effective for profile images and content-based analysis, while Long Short-Term Memory networks capture temporal and sequential behaviour. Transformer-based models such as BERT are applied for textual understanding, enabling improved detection of deceptive content and profile descriptions. When it comes to flagging spam and malicious users, deep learning is highly effective because it can easily parse through huge amounts of data to uncover subtle, hidden patterns that simpler models miss [18].Convolutional neural networks have been explored for fake profile classification using nonlinear activation mechanisms to enhance detection accuracy [19].

D. Graph-Oriented Methods

Graph-oriented methods examine relationships among users through follower networks, interaction graphs, and community structures. These approaches are useful for identifying coordinated fake accounts, bot clusters, and suspicious connectivity behaviour. Graph Neural Networks and embedding techniques

DOI: https://doi.org/10.5281/zenodo.20832761

further enhance detection by learning relational patterns that are difficult to capture through isolated profile features.

Graph-oriented methods analyze social network structures to detect anomalous connectivity patterns and coordinated behaviours among users, enabling effective identification of fake profiles [6, 20]. By utilizing graph-based learning frameworks like Graph Neural Networks (GNNs), these methods can map out relational connections that standard algorithms typically fail to capture [6].

E. Hybrid Detection Methods

Hybrid detection methods integrate machine learning, deep learning and graph-based techniques into a unified framework. This combination improves robustness by leveraging multiple feature sources such as profile attributes, textual content, behavioural signals, and network connections. Hybrid frameworks prove exceptionally robust when tracking down sophisticated fraudulent accounts that operate across various social media platforms. For instance, hybrid deep learning models like DeeProBot integrate diverse neural architectures to boost classification accuracy, analysing a combination of user profile attributes and behavioral tendencies to catch anomalies that single models might skip [4].Advanced hybrid models integrating CNN and BiLSTM architectures have illustrated enhanced output by capturing both structural arrangements and chronological sequences of user behaviour [15].

IV.Tools and Implementation Environment

The proposed study can be implemented using open-source tools commonly used in data-driven security research. Python 3.x is used as the primary programming language with Scikit-learn, TensorFlow, and PyTorch for model development. NetworkX supports graph analysis, while Pandas and NumPy are used for pre-processing. Matplotlib and Seaborn assist in result visualization, and GPU-enabled hardware accelerates deep learning experiments.

The experimental workflow follows a multi-stage pipeline consisting of data collection, preprocessing, feature extraction, model training, and performance evaluation. Datasets from multiple social media platforms are normalized into a common structure to enable cross-platform analysis. Detection outputs are transformed into risk scores, and the system is tested under different attack scenarios to assess accuracy, scalability, and adaptability over time. Performance evaluation techniques such as multi-criteria and fuzzy logic-based approaches can be useful for assessing detection models under varying conditions [22].

V.Analysis and Discussion

A. Comparative Analysis of Detection Techniques

Broadly speaking, the methodologies used to flag fake profiles can be categorized into four primary domains: conventional machine learning, deep learning architectures, graph-based techniques, and hybrid frameworks.

- Traditional Machine Learning methods typically rely on handcrafted metadata and behavioral features. They are computationally efficient and scalable but often struggle to detect sophisticated coordinated activity.
- Deep learning models, particularly CNNs and LSTMs, have proven highly effective at analyzing text and multimedia data. While these architectures generally deliver superior accuracy by mastering complex feature extraction and pattern recognition across multiple domains, their success hinges on massive datasets and substantial computational power [32].
- Graph-Based Methods examine relational structures and community behavior, making them well-suited for detecting coordinated networks.
- Hybrid Models integrate multimodal inputs (text, metadata, and graph structures) and typically achieve the highest reported accuracy [28,33,37] though at the cost of increased system complexity.

Overall, graph-based and hybrid models appear most suitable for cross-platform detection because they capture relational dependencies rather than relying solely on platform-specific features. However, their deployment requires careful engineering to manage computational overhead and real-time processing constraints.

Table 1 provides a practical comparison of commonly used fake profile detection approaches based on deployment-relevant criteria.

DOI: https://doi.org/10.5281/zenodo.20832761

Table 1: Comparison of Fake Profile Detection Methods

Author & Year	Problem	Method	Tools / Techniques	Dataset	Accuracy	Advantages	Limitations
Vishwas Chakranarayan et al. (2025) [38]	Fake account detection	AI-driven multi-modal deep learning (RoBERTa, ConvNeXt, Hetero-GAT)	RoBERTa, ConvNeXt, Hetero-GAT, Deep Learning	FakeNewsNet, Twitter Bot datasets	98.9%	High accuracy, multimodal analysis, robust detection.	Limited dataset diversity, weak against adaptive adversarial attacks, low interpretability.
Dehkordi & Zehmakan et al. (2025)[6]	Fake account detection	Graph-based detection (random walk, GNN, ML models)	Random Walk Algorithms, Belief Propagation, SVM, Random Forest, Node2Vec, Graph2Vec, GCN, GAT, GraphSAGE	Cresci-15, TwiBot-20, TwiBot-22, synthetic graph datasets	High (GNN performs best)	Strong relational analysis, effective for network-based detection.	Limited labeled data, privacy issues, noisy datasets.
Sharma & Singh et al.(2025)[31]	Fake profile detection	Deep learning models (CNN, RNN, BERT, GNN)	CNN, RNN, LSTM, BERT, GNN	Twitter, Facebook, Instagram datasets	97–99%	High accuracy, adaptable DL models.	Poor cross-platform generalization, data imbalance, high computation cost.
Babu et al. (2025) [42]	Fake profile detection on social networking websites to reduce identity theft, misinformation, and security breaches	Comparative machine learning framework using Random Forest, SVM, and Neural Networks with preprocessing and feature engineering	Random Forest, SVM, Neural Network, TF-IDF Vectorization, Label Encoding, Streamlit	User profile dataset in CSV format with features such as screen_name, verification status, statuses_count, followers_count, friends_count, favourites_count	Random Forest: 99.47%, Neural Network: 91.54%, SVM: 85.30%	Higher accuracy with Random Forest; interactive Streamlit application; supports visualization and model comparison; handles structured + textual features.	Dataset source not clearly specified; limited feature diversity; no cross-platform validation; potential overfitting/generalization issues; no adversarial robustness testing.
Prashant Kumar Shukla et al.(2025) [37]	Fraudulent account detection	Hybrid deep transformer model(deep learning architecture)	Temporal Convolutional Network (TCN),Generative Adversarial Network (GAN),Seagull Optimization Algorithm(SOA) (TCN-GAN-SOA framework)	(Cresci-2017 and TwiBot-22)	95% on Cresci-2017 and 94% on TwiBot-22	Allows detecting fraudulent accounts in real time,incorporation of an attention mechanism brings in interpretability.	The analysis was limited to the Twitter datasets (Cresci 2017 and TwiBot-22) and Framework depends mostly on behavioural and textual data.
Jun Li et al. (2024)[40]	Multimodal fake user detection.	MAPM deep learning framework	BERT, multimodal DL	Weibo dataset, BERT text dataset	+27% improvement	Strong multimodal learning, improved performance.	Limited to Chinese datasets (low generalization)
Habib et al. (2024)[25]	Fake profile & content detection.	Hybrid ML & DL (GAN, SVM, CNN, XGBoost)	GAN, SVM, CNN, XGBoost, Random Forest	Twitter, Instagram, StyleGAN, PHEME, GossipCop, Yelp	Up to 99.6%	high accuracy, works across domains, hybrid strength.	Struggles with context/sarcasm, evolving attacks, needs human validation.

DOI: https://doi.org/10.5281/zenodo.20832761

Aditya & Mohanty et al. (2024)[39]	Cross-platform fake profile detection using multimodal data.	Deep-transfer learning + multimodal CNN with optimization	CNN (binary cascaded), Word2Vec, Grey Wolf Optimizer (GWO), Elephant Herding Optimizer (EHO), Fourier Transform, Cosine Transform, Gabor Transform, Wavelet Transform, TensorFlow 2.4, Keras, Python 3.8	Combined datasets (650K samples: Instagram, Facebook, Kaggle, Zenodo)	93.5% (+8.3% improvement)	Multimodal (text, image, video, behavior), real-time detection, reduced redundancy, scalable.	High computational cost, optimization overhead, data imbalance, transfer learning limitations.
Stefanos Chelas et al. (2024)[9]	Detecting fake Instagram accounts to prevent fraud, spam, misinformation, and manipulation of social influence.	Machine learning–based classification with feature engineering using multiple models (Random Forest, Decision Trees, Logistic Regression, MLP, KNN, SVM, Gaussian NB)	Random Forest, Decision Trees, Logistic Regression, MLP, KNN, SVM, Gaussian Naïve Bayes; Feature Engineering (followers/following ratio, etc.)	Combined dataset from InstaFake dataset and Instagram Fake Spammer Genuine Accounts dataset (1890 accounts: 1342 real, 548 fake, 9 features)	97.7% (Random Forest best performance)	High accuracy even with small dataset; effective feature engineering improves performance; multiple model comparison; good precision/recall stability.	Limited dataset size; heavy preprocessing required; data acquisition and privacy constraints; dataset bias and noise; lack of control over social media data; ethical concerns in data usage.
Dharmaraj R. Patil et al.(2024)[41]	Detecting Fake Social Media Profiles.	majority voting approach using machine learning.	Decision Trees, XGBoost,Random Forest, Extra Trees, Logistic Regression,Ada Boost, K-Nearest Neighbors and Majority Voting approach.	MIB dataset contains 3474 authentic accounts, 3351 fraud accounts.	99.12% (Majority Voting approach)	High Performance: Achieves an exceptional 99.12% accuracy, precision, and F1-score .Robust Ensemble: Outperforms individual classifiers by using a majority voting mechanism to synthesize multiple algorithms.	Data Restriction: Focuses exclusively on numeric data,Feature Dependency: Relies on a limited set of 8 specific observable characteristics.

Table 2 : presents a comparative analysis of major fake profile detection approaches based on important operational and analytical features. hybrid and graph-based methods demonstrate stronger cross-platform capabilities, while traditional machine learning techniques offer better explainability and real-time applicability.

DOI: https://doi.org/10.5281/zenodo.20832761

Sr, No	Table 2:Comparative analysis of major fake profile detection approaches						
	Detection Method	High Accuracy	Cross-Platform Capability	Explainability	Real-Time Detection	Graph-Analysis	Scalability
1	Machine Learning	✔	□	✔	✔	□	✔
2	Deep Learning	✔	✔	□	□	□	✔
3	Graph-oriented Method	✔	✔	□	□	✔	□
4	Hybrid Detection	✔	✔	✔	□	✔	✔

Research Gaps

A critical review of the existing literature reveals several prominent gaps that remain unresolved in the domain of fraudulent profile detection:

Deficit in Model Generalization: Current detection methodologies are often over-fitted to specific benchmarks, leading to a marked lack of generalized models that can maintain consistent performance when deployed across disparate social networks or unseen datasets.

Fragmented Feature Integration: While individual studies analyze specific data vectors, there remains an insufficient integration of truly multi-modal data—such as structural profile metadata, temporal behavioural patterns, visual/textual content, and network graph features—within a singular, unified framework.

Scalability and Latency Constraints: Existing deep learning architectures heavily prioritize classification accuracy over computational efficiency, resulting in a distinct lack of research focused on the scalability and real-time execution required for live digital ecosystems.

The "Black Box" Interpretability Barrier: There is inadequate emphasis on embedding explainable artificial intelligence (XAI) paradigms into defensive frameworks, leaving models highly opaque and severely limiting user and systemic trust.

Dataset Obsolescence: Evaluation pipelines frequently rely on out dated or monolithic datasets that fail to reflect the evolving, sophisticated tactics of modern automated bots in contemporary online spaces.

Systematically addressing these limitations establishes the core objective of this study, justifying the development of an intelligent, scalable, and fully transparent fake profile detection framework.

B. Real-World Deployment Challenges

Despite promising experimental results, several obstacles limit real-world implementation:

- The Class Imbalance Problem: Because authentic users vastly outnumber fraudulent ones on real-world platforms, models face severe data imbalance. This skew often biases classifiers toward the majority class, resulting in dangerously high false-negative rates where actual fake profiles slip through undetected.
- Adversarial Adaptation: Malicious actors continuously evolve their strategies to evade detection mechanisms.
- Privacy and Legal Constraints: Cross-platform data integration is restricted by regulatory frameworks and platform policies.
- Lack of Explainability: Several high-performing models operate as black boxes, decreasing clarity and credibility in automated moderation systems.

To mitigate these issues, researchers increasingly explore cost-sensitive learning, continuous retraining pipelines, federated learning frameworks, and explainable AI (XAI) techniques.

Table 3 summarizes these challenges and common mitigation strategies. Deployment challenges and mitigation strategies

Challenge	Impact	Common Mitigation
Data imbalance	False negatives	Resampling, cost-sensitive learning
Adversarial behavior	Model degradation	Continuous retraining
Privacy restrictions	Limited data	Federated learning

DOI: https://doi.org/10.5281/zenodo.20832761

Lack of explainability	**Low trust**	**XAI techniques**

VI.Conclusion

Addressing the global challenge of fake social media accounts requires a multi-faceted approach that spans technological, social, and security dimensions. This research mapped the current defensive landscape by reviewing the strengths and limitations of traditional, deep learning, graph-based, and hybrid methodologies.The evidence highlights a critical trade-off in the field today. Classical machine learning provides highly interpretable and computationally light baselines, but falls short against modern, sophisticated threats. Conversely, hybrid and deep learning frameworks achieve top-tier performance by parsing complex user and multimedia data, though they remain data-hungry and resource-intensive. For fake profile detection to scale effectively against evolving online threats, subsequent research must bridge this gap, prioritizing the development of lightweight yet highly robust hybrid systems.

Future studies should prioritize the integration of explainable AI techniques. Doing so will transform these models from "black boxes" into transparent systems, ultimately fostering deeper trust in automated fake profile detection. Expanding cross-platform deployment and domain adaptation will further strengthen the scalability and real-world applicability of fake profile detection systems.

VII.References

[1] G. Meiselwitz, Ed., Social Computing and Social Media: Design, Ethics, User Behavior, and Social Network Analysis, Lecture Notes in Computer Science, vol. 12194. Cham, Switzerland: Springer, 2020

[2] Y. Hu, L. Manikonda, and S. Kambhampati, “What is Instagram: A first analysis of Instagram photo content and user types,” in Proc. Int. AAAI Conf. Web and Social Media, 2014, pp. 595–598

[3] B. Ghosh et al., “Techniques to detect fake profiles on social media using new-age algorithms: A survey,” IEEE, Jan. 2024.

[4] K. Hayawi et al., “DeeProBot: A hybrid deep neural network model for social bot detection based on user profile data,” Social Network Analysis and Mining, vol. 12, no. 1, p. 43, Dec. 2022

[5] T. Fagni et al., “TweepFake: About detecting deepfake tweets,” PLOS ONE, vol. 16, no. 5, p. e0251415, May 2021

[6] A. S. Dehkordi and A. N. Zehmakan, “Graph-based fake account detection: A survey,” arXiv:2507.06541, Jul. 2025

[7] J. Ezarfelix, N. Jeffrey, and N. Sari, “Systematic literature review: Instagram fake account detection based on machine learning,” EMACS Journal, vol. 4, no. 1, pp. 25–31, Feb. 2022

[8] W. Dracewicz and M. Sepczuk, “Detecting fake accounts on social media portals—The X portal case study,” Sensors, Jun. 2024.

[9] S. Chelas, G. Routis, and I. Roussaki, “Detection of fake Instagram accounts via machine learning techniques,” Computers, vol. 13, no. 11, p. 296, Nov. 2024

[10] E. N. Araka, “A hybrid machine learning model for detection of fake profile accounts on social media networks,” International Journal of Engineering Research and Technology, vol. 13, no. 11, Nov. 2024.

[11] N. Alharbi et al., “Countering social media cybercrime using deep learning: Instagram fake accounts detection,” Sensors, Oct. 2024.

[12] D. G. Shekar et al., “Fake profile detection using deep learning algorithm,” International Research Journal of Engineering and Technology, vol. 11, no. 3, Mar. 2024.

[13] R. Khore et al., “Twitter fake profile detection using deep learning,” International Journal of Research Publication and Reviews, vol. 5, no. 5, May 2024.

[14] A. Agravat et al., “Fake social media profile detection and reporting using machine learning,” International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 5, Mar. 2024.

[15] S. K. Badodia and H. Makwana, “Fake profile detection on social media using hybrid 2D CNN and AES-BiLSTM with network analysis,” Journal of Information Systems Engineering and Management, Feb. 2025.

[16] K. R. Purba, D. Asirvatham, and R. K. Murugesan, “Classification of Instagram fake users using supervised machine learning algorithms,” International Journal of Electrical and Computer Engineering, vol. 10, no. 3, pp. 2763–2772, Jun. 2020

[17] S. R. Sahoo, “Fake profile detection in multimedia big data on online social networks,”

DOI: https://doi.org/10.5281/zenodo.20832761

International Journal of Information and Computer Security, vol. 12, nos. 2–3, 2020.
[18] Z. Alom, B. Carminati, and E. Ferrari, "A deep learning model for Twitter spam detection," Online Social Networks and Media, vol. 18, p. 100079, Jul. 2020
[19] P. Wanda, "RunMax: Fake profile classification using novel nonlinear activation in CNN," Social Network Analysis and Mining, Oct. 2022
[20] M. BalaAnand et al., "An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter," Journal of Supercomputing, vol. 75, no. 9, pp. 6085–6105, Sep. 2019
[21] K. Kaushik et al., "A novel machine learning-based framework for detecting fake Instagram profiles," Concurrency and Computation: Practice and Experience, vol. 34, no. 28, p. e7349, Dec. 2022
[22] S. A. Khan et al., "A novel fuzzy-logic-based multi-criteria metric for performance evaluation of spam email detection algorithms," Applied Sciences, vol. 12, no. 14, p. 7043, Jul. 2022
[23] P. Chakraborty et al., "Fake profile detection using machine learning techniques," Journal of Computer and Communications, vol. 10, no. 10, pp. 74–87, 2022
[24] F. C. Akyon and E. Kalfaoglu, "Instagram fake and automated account detection," in Proc. Innovations in Intelligent Systems and Applications Conf., Oct. 2019, pp. 1–7
[25] S. Khaled, N. El-Tazi, and H. M. O. Mokhtar, "Detecting fake accounts on social media," in Proc. IEEE Int. Conf. Big Data, Seattle, WA, USA, Dec. 2018, pp. 3672–3681
[26] K. Harish et al., "Fake profile detection using machine learning," International Journal of Scientific Research in Science, Engineering and Technology, vol. 10, no. 2, pp. 719–725, Apr. 2023.
[27] J. A. Roberts and M. E. David, "Instagram and TikTok flow states and their association with psychological well-being," Cyberpsychology, Behavior, and Social Networking, vol. 26, no. 2, pp. 80–89, Feb. 2023
[28] R. Bhosale and V. Mane, "A hybrid model for detecting fake profiles in online social networks: Enhancing user trust," Journal of Information Systems Engineering and Management, Jan. 2025.
[29] M. Shivaleela et al., "AI-driven fake profile detection on social media," International Journal of Creative Research Thoughts (IJCRT), vol. 13, no. 5, May 2025.
[30] B. Goyal et al., "Instagram fake profile detection using an ensemble learning method," Scientific Reports, vol. 15, no. 1, p. 26464, Jul. 2025
[31] D. Sharma and N. Singh, "A review of deep learning approaches for fake profile detection on social networking sites," International Journal of Scientific Research in Science, Engineering and Technology, vol. 12, no. 4, pp. 432–445, Aug. 2025
[32] M. Trigka and E. Dritsas, "A comprehensive survey of deep learning approaches in image processing," Sensors, vol. 25, no. 2, p. 531, Jan. 2025
[33] S. R. Sahoo and B. B. Gupta, "Hybrid approach for detection of malicious profiles in Twitter," Computers and Electrical Engineering, vol. 76, pp. 65–81, Jun. 2019
[34] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," IEEE Access, vol. 7, pp. 53040–53065, 2019.
[35] A. Romanov et al., "Detection of fake profiles in social media: A literature review," in Proc. WEBIST, 2017.
[36] M. Dewis and T. Viana, "Phish responder: A hybrid machine learning approach to detect phishing and spam emails," Applied System Innovation, vol. 5, no. 4, p. 73, Jul. 2022
[37] P. K. Shukla, B. D. Veerasamy, N. Alduaiji, S. R. Addula, A. Pandey, and P. K. Shukla, "Fraudulent account detection in social media using hybrid deep transformer model and hyperparameter optimization," Scientific Reports, vol. 15, no. 1, p. 38447, Nov. 2025.
[38] V. Chakranarayan, F. Hussain, F. A. Jaber, R. J. Shaker, and A. Rizwan, "Safeguarding Brand and Platform Credibility Through AI-Based Multi-Model Fake Profile Detection," Future Internet, vol. 17, no. 9, p. 391, Aug. 2025.
[39] B. L. V. S. Aditya and S. N. Mohanty, "Heterogenous Social Media Analysis for Efficient Deep Learning Fake-Profile Identification," IEEE Access, vol. 11, pp. 99339–99351, 2023.
[40] J. Li, W. Jiang, J. Zhang, Y. Shao, and W. Zhu, "Fake User Detection Based on Multi-Model Joint Representation," Information, vol. 15, no. 5, p. 266, May 2024.
[41]Patil, D. R., Pattewar, T. M., Punjabi, V. D. & Pardeshi, S. M. Detecting fake social media profiles using the majority voting approach.EAI

DOI: https://doi.org/10.5281/zenodo.20832761

Endorsed Trans. Scalable Inform. Syst. 11 (3), 1 (2024).
[42]A. Babu, F. Shaik, K. Sri, M. Keerthi, S. Sree, and K. Nandini, “Fake Profile Detection on Social Networking Websites Using Machine Learning:,” in Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies, Nagapattinam, India: SCITEPRESS - Science and Technology Publications, 2025.

DOI: https://doi.org/10.5281/zenodo.20832761