International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

Available Online at :www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET2512523

A Review of Deep Learning Approaches for Fake Profile Detection on Social Networking Sites

Divya Sharma[1], Dr. Nirupma Singh[2]

[1]Research Scholar, School of Engineering & Technology, Career Point University, Kota, Rajasthan, India
[2]Assistant Professor, School of Engineering & Technology, Career Point University, Kota, Rajasthan, India

ARTICLEINFO

Article History:

Accepted: 05 July 2025
Published: 12 August 2025

Publication Issue :
Volume 12, Issue 4
July-August-2025
Page Number :
432-445

ABSTRACT

Social networking sites, now with thousands in existence, have ushered in a revolution in digital communication, while also giving rise to serious security threats like fake profile creation and online impersonations. The perpetrators engaged in these deceitful acts use them for cyberbullying, spreading misinformation, and identity theft, among other evils. Traditional detection methods relying on rule-based systems and shallow, machine learning algorithms have had modest success at best against the increasing complexity of fake profiles. Deep learning approaches have emerged in recent years as powerful, complementary alternatives capable of modeling highly complex patterns from large-scale, heterogeneous sources of data. This review paper presents an in-depth evaluation of state-of-the-art deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Autoencoders, and Transformer-based architecture in the domain of fake profile detection. The paper also looks at multimodal methods combining textual, image, behavioral, and network-based features to enhance detection accuracy. Challenges tackled herein include class imbalance, data privacy, adversarial evasion, and real-time implementation.

Keywords: Fake profile detection, deep learning, social networking sites, online impersonation, CNN, RNN, transformers, multimodal analysis, cybersecurity, identity theft.

INTRODUCTION

The development of social media websites like Facebook, Twitter, Instagram, and LinkedIn has changed the way people all over the world communicate, interact, and express themselves, to a level previously unprecedented. They have now become an indispensable part of life and are deeply involved in interpersonal relationships, company promotions, political discussions and social activism. jpg However, with their fruitful applications and services, social networks have also attracted malicious activities [1]. One of the most predominant threats is profile fabrication and online impersonation, which significantly endangers privacy, trust, and the legitimacy of web communication. Fake profiles are user accounts that give false information about the profile holder, and this false information could be just minimal (partially false) or maximal (fully false) [2]. These profiles are deployed across a host of malicious activities, such as ID fraud, cyber-bullying, phishing attacks, fake news dissemination, financial scams and political manipulation. Online impersonation per se is a more targeted form of deceit which uses simulated real individuals or organizations for the purpose of deceiving their victim(s) into the belief that they are a real entity, pretending to be this entity, and then attracting victims' interests/sympathy/money. Their impact also extends beyond directly affected persons to also include personal and reputational harm, social and political turmoil [3]. It seems that identifying fake personas and imposters has turned out to be a priority for platform providers and information security professionals. Traditional approaches tested were rule based systems, manual moderation, blacklists and traditional machine learning models based on handcrafted features. Although it was a successful ways, it become less efficient to stand out against dynamic and sophisticated attacks. Attackers frequently make use of the deficiency of static rules and of the weak learning model by imitating the behavior of genuine users and evading the ordinary detection method [4].

To address these issues, deep learning has emerged as a potential solutions to fake profile detection [4]. Deep learning models are trained to learn hierarchical representations of raw, unstructured data e.g., text, images, network graphs, in contrast to classical approaches. This can automatically infer complex features and relationships that are hard to describe in manually. Deep learning also enables the creation of end-to-end models for the processing of multimodal information (behavioral patterns, linguistic and visual cues), resulting in more efficient and effective detection systems [5].

In this paper we strive to offer an extensive survey of recent developments in deep learning techniques for the detection of fake profiles and online impersonation in social network sites. Its focus is on exploring a few different deep learning architectures (CNNs, RNNs, “transformer” models, and GNNs) and hybrid approaches that combine multiple modalities. The datasets, evaluation metrics, and challenges typical to this area are also addressed by the survey. In summarizing the existing work and pointing out open issues, our paper aims at directing the future works towards the advancement of intelligent, scalable, as well as privacy-preserving detection systems for protecting online social networks.

A. Objectives of the Review

The main purpose of the present review paper is:

- To investigate the increasing menace of fake identity and online mimicry at social media sites.
- To survey the most recent deep learning methods exploited in detecting fake profiles, including CNNs, RNNs, Transformers and GNNs.
- To study the datasets, features, evaluation metrics used in deep learning-based detection methods.
- To recognize the major challenges in this area, propose potential research directions, and enable better and scalable solutions.

To do so, the paper will act as a reference for researchers, cybersecurity analysts and security professionals looking for efficient deep learning techniques for overcoming fake profiles and online impersonation, which will make these social networking platforms more secure and trustworthy.

B. Overview of Deep Learning Techniques

Deep learning is a branch of machine learning that utilizes multi-layers of artificial neural networks to automatically learn data representations and patterns from large and complex data sets [6]. In contrast to most previous machine learning models, which are characterized by the need for extensive manual feature engineering, deep learning models can automatically learn to extract high-level features from raw data without any prior handcrafted features, ranging from raw text, image, audio to graph structures. This makes them particularly suitable for unstructured or semi-structured data, which is often found on social media sources. Deep learning in the domain of fake profile detection as explained in [6], the main advantage of deep learning is being able to learn complex and non-linear patterns of behavior, language, and interaction that are not generally captured by rule based or shallow models. The Convolutional Neural Network Among the most popular models are Convolutional Neural Networks, which were initially created for image recognition. Besides, in detecting fake profiles, CNN is used to analyze the visual characteristics from profile pictures, posts layouts and even stylometric patterns of account. Its capacity to capture local features via convolutional layers and aggregate them hierarchically makes it powerful in anomaly detection on profile pictures and text structure which might imply the presence of a fake or a bot account [7]. RNNs [8] in which more sophisticated architecture, namely, Long Short-Term Memory (LSTM) networks, can be included, are especially well suited to sequential data. These models are applied to time-series data in the forms of posting rate, user activity logs, or sequences of posts and comments. Because LSTMs capture long term dependencies in temporal sequences, such networks can learn anomaly patterns that relate to bot accounts or trends telling fake users to follow. . For example, a constant posting 24/7 can be regarded suspicious by an RNN model [8]. Another important class of deep learning model is the Autoencoders which are commonly employed in unsupervised learning and anomaly detection. Autoencoders learn a condensed representation of input and then try to homeomorphicly transform the original input. In the context of fake profile detection, they are useful for outlier detection (i.e., for recognizing profiles significantly deviated from the normal user behavior). Profiles with a large reconstruction error may suggest anomalous or fraudulent behaviour [10].

Recently, models based on the Transformer [11] architecture, such as BERT (Bidirectional Encoder Representations from Transformers), have made a major impact on the NLP community. These models are based on self-attention mechanisms to capture context in text. On the fake profile detection front, different forms of BERT have also been used to study the user bio, post text, comments, messages, etc. In combination with natural language approaches, Context classifiers can be leveraged to detect impersonation, spam or malicious content more effectively than common text classifiers [11]. A related emerging deep learning paradigm that has relevance to social networks is Graph Neural Networks (GNNs). Graph-Inspired Representation of Social Media Systems Social media systems (Facebook, Twitter, Instagram) are naturally graph-structured, where users are represented as nodes and the interactions (like, follow, comment) as edges. GNNs are designed to capture features of individual actors as well as relations among them from graph data. GNNs are especially well suited in determining suspicious communities, detecting sybil attacks, and learning network-level anomalies. For instance, graph-based analysis could be used to identify artificial contacts, created with lightweight or otherwise suspicious profile images [12].

Deep learning models are automatic and more flexible as compared to ML models like Support Vector Machines, Decision Trees or Logistic Regression. Traditional models need to be furnished with domain-specific feature engineering while impeding high dimensional or unstructured data. Deep learning, however, lessens the dependence on handcrafted features and can generalize more effectively with larger datasets. However, it has the higher computational cost and the requirement of sufficient labeled data for the supervised learning tasks as well [12].

To summarize: deep learning offers a versatile toolkit of techniques and models suitable for recognizing fake accounts and impersonation in social networking sites. Model selection for various types of data, such as visual, textual, behavioural, or relational data, is regularly conducted, and hybrid systems by multiple models generally obtain a better result.

C. Feature Categories Used in Detection

Deep learning-based fake profiles and online impersonation detection [13] mainly draw on the existence and choice of informative features. These features are used as input to deep learning models and allow the system to differentiate between legitimate and fake user accounts. [13]. Typically, researchers classify them into five main categories: profile-based, behavioral, content-based, network-based and multimodal ones. The categorization reflects various user behaviors and identifications to present a more integrated view of online interactions [14]. Profile-based attributes refer to static characteristics of a user account [15]. They are the username, profile picture, bio, creation date, whether they have an email/phone number, etc. Fake accounts generally contain inconsistencies between these fields, e.g., machine-generated usernames, stock or GAN-generated profile photos, or empty or nonsensical bio sections. For instance, a deep learning model examining visual data can sense abnormalities of profile pictures via convolutional neural networks (CNNs) and then identify repetitive or spam-liked patterns in bios via NLP (natural language processing) models.

Behavioral [16] are able to model the dynamics of the users across the platform. These may be such things as frequency of posts, when they post, number of likes or shares of posts, commenting habits, and engagement rates. Fake or bot accounts tend to demonstrate irregular behaviors like posting regularly regardless of time zone, sending exceptionally high or low levels of engagement and heatwave like bursts of activity. Recurrent neural networks (RNNs) and LSTM models are especially appropriate for capturing such temporal patterns, and they help to detect users that behave abnormally compared to regular behavior over time [16].

Features based on content [17] examine text and visual data provided by the user. These may involve the vocabulary of the posts, attacking intent and sentiment and length of posts among other factors such as to use of hash tags and similarity to known spam or phishing content. Sophisticated NLP models like BERT and RoBERTa can use this training data to learn to detect patterns corresponding to identifying fake accounts, such as the text being overtly-promotional, containing repeated messages between accounts, or linking to harmful sites outside the company site. Sentiment analysis might also expose abnormal sentiment patterns that suggest coordinated bot campaigns or impersonation attacks [17].

The network-based features [18] utilize the social ties to model the structure of the users. The graphs of online social networks are subgraphs of the entire network whose nodes are users and whose edges represent various kinds of interaction between them (friendship, be/provided by mentions, followers, and so on). Fake accounts tend to lack strong connections, have unusual follower-to-following ratios, or are within abnormally dense neighborhoods all jockeying to seem legitimate. Graph Neural Networks (GNNs) and other graphbased learning models can be used to model these relationships and detect anomalies in the connectivity, centrality, and community structure indicative of fraud. Lastly, multimodal fusions are combinations of two or more types of features presented above, aiming to enhance the detection rate. In a hybrid model, for example, CNNs may be used to analyze profile pictures, LSTMs may be used to process re-post behavior, transformers may be used to analyze the textual content, with the output being fed into an integrated decision-making system. Combinations of modalities are particularly effective for countering advanced fake profiles that manage to impersonate natural users along one dimension but falter when scrutinized across many dimensions.

In conclusion, feature engineering is essential for deep learning models in fake profile detection to be effective. Although each category makes its own contribution to detection, the integration of all categories will allow models to capture more comprehensively the multifariousness of online deception.

D. Deep Learning Approaches for Fake Profile Detection

With the proliferation and scale of social networking platforms, deep learning for automated fake profile and online impersonation has been recognized as a powerful tool. A vast number of different architectural models such as convolutional and recurrent neural networks, transformers and graphs have been used to learn complex patterns in user data. Such model types are tailored to different types of data and detection tasks and, to a large extent, mixed model types are applied to reach the best detection performance. CNN Convolutional Neural Network (CNN) [19] is a type of neural network which is broadly used for computer vision tasks, and it has been taken into consideration for fake profile detection in particular in profile images and visual content. Fake accounts generally rely on stolen, edited, or images obtained from GANs that look like real to human eyes, but have visual artifacts, which can be recognized using CNNs. For example, authors have trained CNN models to distinguish between real images and generated images by detecting artifacts at the pixel level. CNNs can even be used for text data – feeding word embeddings as input matrices, the model will be capable of capturing local semantic patterns in user bios/posts.

The recent development of Recurrent Neural Networks (RNNs) [20] and Long Short-Term Memory networks based on them [21] have been proven to be effective in modelling the temporal behavior in sequential data. These models have been applied to track queuing of posting order, activity sequences and engagement across time. Fake profiles or bots usually exhibit abnormal interaction rhythms (e.g., both too frequent and happened at strange time and huge number of burst in a short period of time), which can be captured by LSTM through learning temporal dynamics. For instance, a formulation using RNN to monitor time and action types of user interactions over time and mark accounts that exhibit repetitive or automated behaviors [20].

CNN-RNN hybrid models [21] capture the advantages of multi-modal types. A typical solution is to use CNN to model visual or textural representations and feed them into RNN to consider temporality or sequences of user behavior. These hybrid models are particularly strong in complex scenarios, where content (images, text) and behavior (activity logs) should be analyzed together. For instance, a model that combines visual authenticity of a profile image and time-series activities of posts can decide the likelihood of an account being fake. Transformer-based models [21] including BERT (Bidirectional Encoder Representations from Transformers) has changed the natural language processing (NLP) landscape, and these models are starting to be used in fake profile detection. These models rely on self-attention mechanisms to capture local context spans in sentences. models -spectrogram and model) to achieve high accuracy in predicting the user bio, comments, direct message, and post caption as opposed to all BERT- based ones. They can also identify any attempted impression by the writing style, emotional tone and language, among others, inconsistent with the user’s known linguistic profile. Transformers also perform well at cross-lingual detection, an important requirement of multilingual social media.

Graph Neural Networks (GNNs) [22] are emerging for mining the relational data structure of social networks. Social networks are naturally modeled as graphs, where a user is represented by a node and an edge captures the interaction (eg., follow, like, mention). GNNs learn and extract embeddings which not only consider the feature of nodes but also their topological context, and allow to detect a variety of suspicious patterns such as lilypads, dense bot clusters and unusual community structures. Fake and genuine users may have varying shallow or highly artificial connectivity offset, which can be identified by GNNs, for instance.

Finally, the selection of the appropriate deep learning architecture for fake profile detection depends mostly on how data are presented. Text, image, behavior, and social network analysis, deep learning can be applied to find malicious accounts and impersonators with more accuracy and scalability for different attempts.

METHODOLOGY

The purpose of this review paper is to provide a comprehensive survey on the latest research in the domain of deep learning techniques used for detecting fake profiles and online impersonation on social network platforms. A systematic literature review approach was incorporated to limit the pool of articles to provide high quality, relevant, recent contributions. The methods include a systematic search method, inclusion/exclusion criteria focusing on this specific view and a multi-pass sift process to assure the academic and relevance quality of the data that were included. The methodology adopted is discussed in the following subsections:

A. Databases/Resources

The review on literature presented in this article was based on an operated sample list which contains notable content from authoritative academic retrieval resources that are well accepted in the publication of high-quality and peer-reviewed research works on artificial intelligence, cybersecurity, deep learning and social network analysis. These include:

- IEEE Xplore
- SpringerLink
- Elsevier (ScienceDirect)
- Wiley Online Library
- Google Scholar
- ACM Digital Library

In addition to these databases, we also searched proceedings of renowned AI conferences namely NeurIPS, ICML, and AAAI to find the recent trends for deep learning based user profiling, fake account detection, and bot activity analysis.

B. Inclusion Criteria

The preserved significance and quality of studies proved the validity of inclusion criterion. We did not consider pre-2019 research, however a handful of seminal papers published in earlier years were included should some degree of historical grounding prove essential. The following criteria were used to review the literature:

- Topic The paper should address fake profile detection, bot detection or impersonation on social networks.
- Deep Learning Exciting: Only works employing deep learning such as CNNs, RNNs, LSTMs, etc… • Data Format: Research relying on profile data, user behaviors, content (textual) analysis, or network graph structures were considered predominately.
- Exclusion: Studies that used only standard machine learning (like Naïve Bayes or Decision Trees) or studies not aligned with topic social media security were excluded.

C. Keywords Used

A keyword search strategy was applied, with selected terms, often in conjunction with Boolean operators (AND, OR) to retrieve targeted and relevant literature. The main keywords included:

- Discovering fake profiles in social networks

- Deep learning for social network security
- CNN、RNN、GNN etc..
- Online impersonation detection with AI
- Transformer models for social network analysis
- Fake account identification using graph neural network
- Deep learning in social media for fraud detection

These keywords facilitated the thorough but focused coverage with respect to the relevant literature of the review.

D. Stepwise Selection Process

In order to investigate the most relevant and high-impact study, a stepwise and organized election process has been pursued.

- Collection: The original search identified over 100 articles through the selected databases with the keywords. Citation search was also conducted to expose recognized references frequently referred in recent research.
- Selection: Reviewing titles and abstracts and further excluding papers that did not meet the inclusion criteria. This yielded 25 articles shortlisted by their content relevance, emphasis on deep learning method and social network application.
- Final Review After assessing for full text, 15 studies of high quality were included in this review. These were state-of-the-art studies in CNN-based image analysis, RNN-based behavioral modeling, transformer-based content analysis, and GNN-based network structure modeling for the identification of fake profiles and online impersonation.

E. State-of-the-Art and Emerging Trends Focus

The chosen literature overall represents the state-of-the-art in deep learning for fake profile detection, and the following new research trends that are emerging, including:

- Explainable AI (XAI): Interpretable detection systems in order to enhance transparency and trust.
- Federated training: Collaborative models between platforms with privacy preservation.
- Multimodal fusion: Combining text, image and graph for enhanced detection. • Cross-Platform Detection May contain the models that can transfer across different social networking platforms.

Adopting a systematic method, this review guarantees a well-rounded and current comprehension of methods, difficulties, and future directions in using deep learning to fight identity fraud and online impersonation in social media.

Table 1: Selection Process of Literature Review

Stage	**Number of Papers**	**Description**
Initial Collection	100+	Articles retrieved using targeted keywords and citation tracking from major academic databases.
Shortlisting	25	Screened for relevance, deep learning focus, and alignment with social media fake profile topics.
Final Review	15	In-depth review of most influential studies using CNN, RNN, BERT, GNN, and hybrid architectures.

Total 100
Shortlisted 25
Final 15

Figure 1. Funnel Diagram for Literature Review

LITERATURE REVIEW

The selected final papers for literature review are as follows,

Winston, L., et al. (2025) [23]: In this paper, we investigate the use of deep learning models, such as RNN, LSTM, CNN, GNN, and transformers, to identify fake accounts through their behaviors such as clickstreams and session information. Experiments on various corpora show that both temporal and relational models are more effective than state-of-the-art models. Main concerns are data imbalance, interpretability and real-time performance, the conclusion being that deep learning is a scalable solution.

Bashaddadh, O., et al. (2025) [24]: This review of 90 papers describes the Transformer-based methods such as BERT being popular among fake news detection systems with almost negligible error, by following PRISMA methodology and the research papers published between the year 2020–2024. GAN-based and multimodal solutions look promising, but challenges still exist in terms of dataset diversity and generalization, model interpretability. The authors call for ethical, multilingual and efficient models in upcoming research and development.

Reddy, K. M. S. M. S. (2025) [25]: The authors propose a machine learning and deep learning-based web system for identifying fake accounts, originaly developed for Instagram. It distinguishes real from fake profiles and is designed as relatively scalable to expand to other platforms, dealing with the limitations of manual detection in anonymous social environments.

Will, I., et al. (2025) [26]: Concentrating on user authentication, this work leverages GNNs to represent user interactions in the form of graphs and detect fake or impostor profiles. The paper covers usages and performance when applying these models, including adversarial attacks, privacy preserving and data skew, with future topic on privacy preserving GNNs..

Wang, Z., et al. (2025) [27]: This study introduces a behavior-based profiling framework, adapts the 5Cs model to a 3Cs strategy and applies stacking ensemble technique. Our approach outperforms the baseline by 9.26% accuracy and 3.69% of AUC in imbalanced dataset, highlighting the efficacy of deep behavioral modeling to curtail manipulative behaviors on social networks.

Habib, A. R. R., et al. (2024c January) [28]: Based on a survey of state-of-the-art machine learning methods, this work explores fake account detection in OSNs that targets profile content, images, and behavioural signals. It challenges the restricted approach of manual identification and seeks existing automated remedies for phishing, misinformation, and manipulation.

Shah, A., et al. (2024) [29]: The paper compares different machine and deep learning models: LSTM is better on large datasets; traditional ones (e.g., XGBoose) work better on smaller data.

Alharbi, N., et al. (2024) [30]: This paper presents an LSTM model to detect fake Instagram accounts, achieving accuracies of 97.42%, 94.21%, and 99.42% in varying datasets. The model exhibits good accuracy and robustness for various platforms, highlighting the universal influence of fake accounts on digital platforms.

Unni, M. V., et al. (2024) [31]: With the hybrid detection system, namely FPD-COADL, with the application of the Coyote Optimization Algorithm and Deep Belief Networks, this study uses the profile features and content information to detect a fake user. Experiments show promising accuracy and scalability, making the proposed method a strong candidate as an automatic solution for trusting the digital worlds.

Deedee, B., et al. [32]ALERGSHERBANA et al.(2024) This work is Our work make use of RF and DCNN to detect fake profiles and stalking on X (Twitter). When processing user metadata, the model achieves 93.89 % accuracy. Designed basing on principle of OOAD and implemented by Python, it can outperform the comparison system in real time anomaly detection and behavior monitoring.

Suryawanshi, J., et al. (2024) [33]: Aiming at fake profiles in recommendation systems, the Voting Ensemble model is proposed which combines QDA, KNN and NBC. It projects over 99% accuracy on the Movielens dataset, and performs precision, recall and MCC advances, consolidating RS reliability towards profile manipulation.

Lubis, A., et al. 28 (2024) [34]: Focusing on Twitter fake accounts, the research compares traditional models and CNNs. CNNs perform better by 86 \% (and with transfer learning reached 93.9 \%). The experiment underscores that CNNs can succeed in spotting misleading low-activity accounts, even if static features are ineffective.

Venkatesh, S. C., et al. (2024) [35]: Introducing a multistage stacked ensemble model for the detection of fake profiles, this employs chi-square feature selection and cost-sensitive learning. It obtained high precision in the Facebook (95%), Instagram (98.2%), and Twitter (81%) datasets, proving that ensembles still scale properly and perform well in practice.

Yigezu, M., et al. (2024, March)[36]: RNN-LSTM models are used for fake news detection using optimized hyperparameters. However, its performance is not so good in multiclass mode because it is influenced by the imbalance of data (0.82 score for binary). Writers suggest balanced datasets to improve the performance of such future multi-label.

George, N., et al. (2024) [37] :By proposing the ransomware detection model based on CNN and RNN layers that synthesizes the system activities to obtain the dynamic behavioral signatures. It dramatically enhances the detection rate and minimizes false positive result with latency showing potential for real-time applications for large-scale cybersecurity applications.

Table 1. Literature Review Findings

Author Name (Year)	**Main Concept**	**Findings**	**Limitations**
Winston, L., et al. (2025)	Behavior-based fake account detection using deep learning	Models like RNN, LSTM, CNN, GNN, and Transformers outperform traditional methods on dynamic data.	Data imbalance, limited interpretability, and real-time application challenges.
Bashaddadh, O., et al. (2025)	Systematic review of fake news detection using ML and DL	Transformer models (e.g., BERT) show up to 99.9% accuracy; multimodal and GAN-based models are promising.	Dataset diversity, model generalizability, and explainability issues.
Reddy, K. M. S. M. S. (2025)	Web-based system for fake profile detection on Instagram	Platform-specific modules accurately classify profiles as real or fake; scalable system architecture.	Limited to Instagram; expansion to other platforms pending.
Will, I., et al. (2025)	GNN-based user verification system	GNN effectively models user relationships for detecting fake accounts and impersonation.	Vulnerable to adversarial attacks; privacy and data imbalance concerns.
Wang, Z., et al.	Behavior profiling using	9.26% accuracy and 3.69% AUC	Imbalanced data; model

Author Name (Year)	Main Concept	Findings	Limitations
(2025)	ensemble learning for social manipulation detection	improvement using 3Cs model and stacking ensemble.	interpretability needs improvement.
Habib, A. R. R., et al. (2024)	Review of fake account detection in OSNs	ML approaches identify bots using content, image, and behavior cues; manual detection is unreliable.	Lack of real-time tools and generalization across platforms.
Shah, A., et al. (2024)	Model performance analysis on varying dataset sizes	LSTM achieves 97% on large datasets; XGBoost performs equally well on smaller datasets.	Effectiveness depends on dataset size and balance.
Alharbi, N., et al. (2024)	LSTM-based detection of fake Instagram profiles	Achieves 97.42%, 94.21%, and 99.42% accuracy across Instagram and Twitter; high F-measure.	Focused on two platforms; generalization across others untested.
Unni, M. V., et al. (2024)	Hybrid detection using COA and Deep Belief Network (FPD-COADL)	High accuracy and scalability in detecting fake profiles using content and behavior.	Custom dataset; limited comparative evaluation.
Deedee, B., et al. (2024)	RF and DCNN-based model for fake profile and stalker detection on Twitter	Achieves 93.89% accuracy; effective in real-time behavioral anomaly detection.	Limited to Twitter; uses specific features that may not generalize.
Suryawanshi, J., et al. (2024)	Ensemble model (VE) for detecting fake profiles in recommendation systems	VE model achieves 99.6% accuracy, outperforming QDA, KNN, and NBC individually.	Tested only on Movielens; lacks diversity in platform data.
Lubis, A., et al. (2024)	CNN-based fake account detection on Twitter	CNN achieves 86%; transfer learning raises it to 93.9%, outperforming traditional methods.	Struggles with low-activity deceptive profiles using static features.
Venkatesh, S. C., et al. (2024)	Stacked ensemble model with feature selection and cost-sensitive learning	Precision scores: Facebook (95%), Instagram (98.2%), Twitter (81%). Outperforms conventional models.	Complex model pipeline; Twitter performance lower.
Yigezu, M., et al. (2024)	Fake news detection using RNN-LSTM with grid search	Binary classification score of 0.82; limited multi-class performance due to unbalanced data.	Poor multi-label performance; recommends data balancing strategies.
George, N., et al. (2024)	DeepCodeLock framework for ransomware detection	CNN-RNN architecture detects complex ransomware behavior	Focused on cybersecurity, not social media;

Author Name (Year)	Main Concept	Findings	Limitations
		with high accuracy and low latency.	applicability in social contexts not tested.

RESEARCH GAPS DISCUSSION

Despite the advances in machine learning and deep models to detect fake users and profiles on social media, there are still some research gaps. One of the key issues is the lack of generalization, as most models are developed to a particular platform or dataset and do not work well for other datasets. Furthermore, the issue of data imbalance remains as fake profiles are under-represented in training data which results in biased predictions, despite measures such as SMOTE and ensemble methods. Some are conducted using synthetic or artificially-manipulated data that fails to capture the natural complexities. It is also difficult to interpret deep learning models: very accurate models such as Transformers and GNNs are often black boxes with poor transparency and trust. Furthermore, the majority of the research is for English-point-of-care data and there is lack of work on language and culturally sensitive computational methods. In-addition, real-time detection is challenging as most of the existing models require too many computational resources to be applied for live monitoring. Lastly, ethical issues -- including user privacy, generation of false positives, and resistance against adversarial attacks -- are often inadequately considered. These shortcomings demonstrate the need for more effective and more interpretable, adaptable, and ethical methods to detect fake profiles.

CONCLUSION

Although much improvement has been made in the aspect of machine learning and deep models that aim to spot fake users and profiles on social media, several research gaps still exist. One of the main reasons is that the models can not be generalized, because they are constructed based on certain database and the way of generating the database. Additionally, data imbalance is also the area of concern even when the fake profiles are underrepresented in the training data leading to the biased predictions, in spite of using tactics like SMOTE, and ensemble methods. Some of which are based on the synthetic or artificially manipulated data which does not represent the complexity of the nature. Deep learning models are also difficult to interpret: even very accurate models like Transformers and GNNs are often black-boxed with bad transparency and trust. In addition, most of the work is restricted to English-point-of-care data and there has been little research on language and culturally sensitive computational approaches. Moreover, real-time detection becomes very difficult because of most of the existing models used in practice requires excessive computational resources to be used in live monitoring. Finally, there is a lack of concern by practitioners about ethical questions, such as user privacy, presence of false-positives, and immunity against adversarial efforts. Such failures call for more robust, interpretable, flexible, and ethical techniques to identify fake profiles.

REFERENCES

[1]. Aditya, B. L., & Mohanty, S. N. (2023). Heterogenous social media analysis for efficient deep learning fake-profile identification. IEEE Access, 11, 99339-99351.

[2]. Amankeldin, D., Kurmangaziyeva, L., Mailybayeva, A., Glazyrina, N., Zhumadillayeva, A., & Karasheva, N. (2023). Deep Neural Network for Detecting Fake Profiles in Social Networks. Computer Systems Science & Engineering, 47(1).

[3]. Ahmad, S., & Tripathi, M. M. (2023). A review article on detection of fake profile on social-media. International Journal of Innovation Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.

[4]. Oulad-Kaddour, M., Haddadou, H., Vilda, C. C., Palacios-Alonso, D., Benatchba, K., & Cabello, E. (2023). Deep learning-based gender classification by training with fake data. IEEE Access, 11, 120766-120779.

[5]. Alsubaei, F. S. (2023). Detection of inappropriate tweets linked to fake accounts on twitter. Applied Sciences, 13(5), 3013.

[6]. Sukanya, L., Aniketh, J., Abhiman Sathwik, E., Sridhar Reddy, M., & Hemanth Kumar, N. (2023). Racism detection using deep learning techniques. In E3S Web of Conferences (Vol. 391, p. 01052). EDP Sciences.

[7]. Bordbar, J., Mohammadrezaei, M., Ardalan, S., & Shiri, M. E. (2023). Detecting fake accounts through generative adversarial network in online social media.

[8]. Aboud, A., Rokbani, N., Mirjalili, S., Hussain, A., Chabchoub, H., & Alimi, A. M. (2023). A quantum beta distributed multi-objective particle swarm optimization algorithm for twitter fake accounts detection.

[9]. Ojugo, A. A., Akazue, M. I., Ejeh, P. O., Odiakaose, C., & Emordi, F. U. (2023). DeGATraMoNN: Deep Learning Memetic Ensemble to Detect Spam Threats via a Content-Based Processing. Kongzhi yu Juece/Control Decis, 38(01), 667-678.

[10]. Siddiqui, F., & Suaib, M. (2023). Enhancing Spammer Fake Profile Detection on Social Media Platforms using Artificial Neural Networks. International Journal of Engineering and Management Research, 13(4), 1-6.

[11]. Kanagavalli, N., & Priya, S. B. (2022). Social networks fake account and fake news identification with reliable deep learning. Intell. Autom. Soft Comput, 33(1), 191-205.

[12]. Suganthi, S. T., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. PeerJ Computer Science, 8, e881.

[13]. Wanda, P. (2022). RunMax: fake profile classification using novel nonlinear activation in CNN. Social Network Analysis and Mining, 12(1), 158.

[14]. Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2022). Are deep learning-generated social media profiles indistinguishable from real profiles?. arXiv preprint arXiv:2209.07214.

[15]. Voitovych, O., Kupershtein, L., & Holovenko, V. (2022). Detection of Fake Accounts in Social Media. Кібербезпека: освіта, наука, техніка. Т. 2,№ 18: 86-98.

[16]. Kadam, N., & Sharma, S. K. (2022). Social media fake profile detection using data mining technique. Journal of Advances in Information Technology Vol, 13(5).

[17]. Benabbou, F., Boukhouima, H., & Sael, N. (2022). Fake accounts detection system based on bidirectional gated recurrent unit neural network. International Journal of Electrical and Computer Engineering (IJECE), 12(3), 3129.

[18]. Ali, A. K., & Abdullah, A. M. (2022). Fake accounts detection on social media using stack ensemble system. International Journal of Electrical and Computer Engineering (IJECE), 12(3), 3013-3022.

[19]. Boahen, E. K., Bouya-Moko, B. E., Qamar, F., & Wang, C. (2022). A deep learning approach to online social network account compromisation. IEEE Transactions on Computational Social Systems, 10(6), 3204-3216.

[20]. Cartwright, B., Frank, R., Weir, G., & Padda, K. (2022). Detecting and responding to hostile

disinformation activities on social media using machine learning and deep neural networks. Neural Computing and Applications, 34(18), 15141-15163.

[21]. Singh, V., Shanmugam, R., & Awasthi, S. (2021). Preventing fake accounts on social media using face recognition based on convolutional neural network. In Sustainable Communication Networks and Application: Proceedings of ICSCN 2020 (pp. 227-241). Springer Singapore.

[22]. Kesharwani, M., Kumari, S., & Niranjan, V. (2021). Detecting fake social media account using deep neural networking. International Research Journal of Engineering and Technology (IRJET), 8(7), 1191-1197.

[23]. Winston, L., Omoseebi, A., & Collines, J. (2025). Deep Learning Models for Behavior-Based Fake Account Detection.

[24]. Bashaddadh, O., Omar, N., Mohd, M., & Khalid, M. N. A. (2025). Machine Learning and Deep Learning Approaches for Fake News Detection: A Systematic Review of Techniques, Challenges, and Advancements. IEEE Access.

[25]. Reddy, K. M. S. M. S. (2025). Spammer Detection and Fake user Identification on Social Networks.

[26]. Will, I., Omoseebi, A., & Jhon, J. (2025). Graph Neural Networks (GNNs) for Social Media Profile Verification.

[27]. Wang, Z., Li, L., He, K., & Zhu, Z. (2025). User Profile Construction Based on High-Dimensional Features Extracted by Stacking Ensemble Learning. Applied Sciences, 15(3), 1224.

[28]. Habib, A. R. R., Akpan, E. E., Ghosh, B., & Dutta, I. K. (2024, January). Techniques to detect fake profiles on social media using the new age algorithms-A Survey. In 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0329-0335). IEEE.

[29]. Shah, A., Varshney, S., & Mehrotra, M. (2024). Detection of fake profiles on online social network platforms: performance evaluation of artificial intelligence techniques. SN Computer Science, 5(5), 489.

[30]. Alharbi, N., Alkalifah, B., Alqarawi, G., & Rassam, M. A. (2024). Countering Social Media Cybercrime Using Deep Learning: Instagram Fake Accounts Detection. Future Internet, 16(10), 367.

[31]. Unni, M. V., Jeevananda, S., Kalapurackal, J. J., & Fatma, S. (2024). Enhancing authenticity and trust in social media: an automated approach for detecting fake profiles. Indonesian Journal of Electrical Engineering and Computer Science, 35(1), 292-300.

[32]. Deedee, B., Onate, T., & Emmah, V. (2024). Fake Profile Detection and Stalking Prediction on X using Random Forest and Deep Convolutional Neural Networks. Journal Press India, 4(1).

[33]. Suryawanshi, J., Abdul, S. M., Lal, R. P., Aramanda, A., Hoque, N., & Yusoff, N. (2024). Enhanced Recommender Systems with the Removal of Fake User Profiles. Procedia Computer Science, 235, 347-360.

[34]. Lubis, A., Prayudani, S., Hamzah, M. L., Lase, Y., Lubis, M., Al-Khowarizmi, A., & Hutagalung, G. (2024). Deep neural networks approach with transfer learning to detect fake accounts social media on Twitter. Indones. J. Electr. Eng. Comput. Sci, 33, 269.

[35]. Venkatesh, S. C., Shaji, S., & Sundaram, B. M. (2024). A fake profile detection model using multistage stacked ensemble classification. Proc Eng Technol Innov, 26, 18-32.

[36]. Yigezu, M., Kolesnikova, O., Sidorov, G., & Gelbukh, A. (2024, March). Habesha@ dravidianlangtech 2024: Detecting fake news

detection in dravidian languages using deep learning. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 156-161).

[37]. George, N., Sham, A., Ajith, T., & Bastos, M. (2024). Forty Thousand Fake Twitter Profiles: A Computational Framework for the Visual Analysis of Social Media Propaganda. Social science computer review, 08944393241269394.