International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

A review on Fake Profile Detection Methods Using Machine Learning Approaches

Deep Shikha
Department of Computer Science and Engineering
Research Scholar, Rajshree Institute of Management and Technology, Bareilly
Dr Ruchin Jain (HOD)
Department of Computer Science and Engineering
Guide, Rajshree Institute of Management and Technology, Bareilly

Abstract
The increasing use of social media and digital platforms has led to a surge in fake profiles, which are often created for malicious purposes such as spreading misinformation, committing fraud, or manipulating public opinion. These fake accounts pose a serious threat to user safety, trust, and platform integrity. Traditional rule-based methods and manual moderation are no longer sufficient to handle the scale and sophistication of these deceptive profiles. As a result, researchers have turned to machine learning (ML) techniques to develop more effective and automated detection systems. This review explores various machine learning approaches used in fake profile detection, including supervised learning models like Decision Trees, Random Forests, and Support Vector Machines, as well as unsupervised techniques such as clustering and anomaly detection. It also highlights the application of deep learning and graph-based models for analyzing complex user behavior and network patterns. Key stages such as feature extraction, model training, and evaluation are discussed, along with the importance of selecting relevant features such as posting frequency, sentiment of messages, and social network metrics. The review also addresses current challenges such as class imbalance, adversarial tactics, and dataset limitations. In conclusion, machine learning offers a powerful and scalable solution for detecting fake profiles, but continuous innovation is needed to stay ahead of evolving threats.

Keywords: Fake profiles, Machine learning, Supervised learning, Anomaly detection, social media security

Introduction

The problem with fake profiles on social networking platforms is that they pose serious threats, like misinformation sowing, phishing attacks, identity theft. In order to address these issues, researchers have more and more begun to leverage machine learning (ML) approaches for

detecting and flagging fake profiles. The traditional rule-based systems fail however as they are unable to accommodate changes in the behavior of the authors of fake accounts. On the other hand, machine learning models learn from data, thus, are more adaptive and are more accurate. Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks have been widely used, but are considered supervised learning techniques. Finally, these models use features extracted from user profiles, e.g.f friend count, post frequency, profile completeness, and sentiment of a post. It has been found that feature selection is very important to improve accuracy. For example, fake profile may have abnormal friend to follower ratio or suddenly got so active. Moreover, natural language processing (NLP) methods can be leveraged to extract the tone and content of the user posts to find deceptive patterns.

Clustering procedures and anomaly detection models alike have demonstrated the potential to find fake profiles when labelled data is limited. The goal of this application is to do user profile grouping and flag outliers that differ from normal behavior. However, recent advancements in graph-based machine learning have brought social graph analysis into the picture where the network is used to model the relationships between users. Graph anomalies can be used by these models to reveal botnets or coordinated sets of fake profiles. In addition, multi-ML technique hybrid models tend to perform better than single model approaches. Although progress has been made in thse areas, problems like class imbalance, limited availability of high-quality data sets and the use of adversarial tactics by attackers still exist. Future research on ensemble models, transfer learning and real time detection systems are developed to overcome these issues. There's also rising interest in integrate explainable AI (XAI) to increase transparency in decision making. Altogether, the use of machine learning has contributed greatly to improving the effectiveness of fake profile detection systems.

Significance of the Study

Therefore, the importance of learning the technique of fake profile detection with the help of machine learning is very important as it contributes towards a safe online presence, building the user's trust and integrity of the online platform. Online fake profiles associated with social media, e-commerce and other Internet communities are a huge threat given the rapid growth of new Internet areas. Like traditional detection methods, static rules or manual review are not adequate anymore considering the dynamics and sophistication of fake profiles. A potentially more powerful alternative is provided by machine learning approaches that automatically learn patterns of usage and detect anomalies in usage. By doing so, these methods help platforms

have better ability to detect fake accounts more efficiently and with less manual cost and response time.

Additionally, the study adds to the body of research in the design of intelligent, scalable, and real time monitoring systems with the ability to accommodate changing threats. Together with machine learning techniques such as supervised classification, unsupervised clustering, and deep learning, researchers can reveal hidden patterns that are hardly traceable by traditional means. As such, this research also offers additional protections to real users from scams, identity theft and even cyberbullying by proactively identifying and removing malicious entities. In a broader sense, the results could be of help in other domains of information, including online banking, dating and job portal websites, that need to foster trust and authenticity. Furthermore, this research also promotes transparency and accountability in automated systems especially in the light of the present-day importance of explainable AI. In the end, this research does not only enhance cybersecurity but also encourages ethical use of artificial intelligence in digital media.

Background and Rise of Fake Profiles

In our digital day and age, there has been an ongoing rise of fake profiles on online platforms due to the growing popularity and reach of social media, e commerce, dating apps as well as professional networking sites. One of the biggest reasons why the fake profiles are formed is to deceive users by impersonating the real person or creating a totally fake identity. Such profiles are used to achieve various malicious purposes at a time of elections or global events, such as to spread misinformation, execute scams, manipulate public opinion and so on.

Thanks to technological innovation, it’s easier than ever to build fake profiles that are realistic, believable, even real to the unsuspecting, with no effort on the part of the bad actor—they just needed an automated bot and a photo generated by AI. With occasionally convincing photos, bios and posts and deep fake technology and generative text models, fake profiles have become difficult to identify manually. Further complicating the issue is the number of users on a large platform which allows it to be easy to hide in a sea of fake account.

Many times, fake profiles leverage the trust dynamic to interact or prey on vulnerable users for personal gain or agendas that could end up being damaging or even disruptive. As they are widely present, online environments are undermined from a credibility point of view for individuals and organisations. The generation of fake profile at sophisticated and massive scale has rendered human moderation and static rule inconceivable. Priority is placed on developing

intelligent, adaptable and automated detection systems, especially those based on machine learning techniques, to tackle the threat to the border landscape which is developing rapidly.

Impact on Social Platforms and Users

In the abundance of fake profiles is a deep and wide spread impact to both social platforms and their users. Fake accounts constitute a direct threat to online safety and privacy for individuals, in particular. A user might be tricked into sharing personal information, giving money away, or even entering into emotional relationships with completely bogus personas, which is known as ‘catfishing’. Due to these experiences, psychological distress, financial loss, and a lack of trust in digital interactions, arise. Fake profiles ruin the ecosystem of the online community on a wider level. The majority of these people use them for scamming, creating misleading impressions about the follower count, and spaming/spreading untruthful content. That negatively affects the user experience, undermining the platform’s trustworthiness. Fake profiles is affecting marketing analytics, creating a fake perception of the brand sentiment and destroying the businesses’ reputation. Furthermore, the dishonesty is also introduced by influencers or businessperson who purchases fake followers to uplift the testimonial of the business and social metrics. Fake profile management is a big problem for platform providers that has to be fixed constantly and isn’t cheap in terms of resources demand and the need of coming up with new ideas every time. Having these accounts causes risks of regulatory scrutiny, legal consequences and a decrease in user engagement. If users think the platform is insecure or fake activity filled, they will limit their time and quit eventually. This is why security fight against the fake profile is not just a matter of security but a pillar to digital platforms' sustainability and growth. However, for the purpose of addressing such challenge effectively, a proactive and scalable approach like using machine learning based detection system is needed.

Need for Automated Detection

Lack of manual moderation and rule-based systems to cope with scale and complexity in modern digital platforms create the need for automation in fake profile detection. With a surging number of online users, the number of fake accounts by malicious people is also exploding. Most of these profiles are run and operated by bots or humans working in concert that can mimic usual user behaviour and also evade the traditional detection methods. However, the time consuming, expensive and impractical manual review processes are not feasible for real time detection across the millions of accounts. Static, rule based systems for things such

as identifying profiles with no photos or too much activity are easily circumvented by advanced fake account creators looking to meet guidelines on a platform. In the context of large volumes of data, machine learning offers a scalable and adaptive solution to learn from such volumes to look for the subtle patterns and anomalies that might shed light on the fake behavior. With the help of machine learning techniques, there are automated detections where platforms can assess lots of features, from network behavior, until post content, activity timing, and interaction patterns, with big efficiency and accuracy. These systems can learn from historical data and be updated continuously to evolve with the new tactics employed by the attackers. Additionally, real time detection allows fake profiles to be addressed before any damage is done, improving user trust and safety. the biggest threat and rapidly changing nature of fake profiles demand a shift to intelligent, automatic detection systems using the power of machine learning for effective and proactive defences against them.

Literature review

Goyal, B. et al (2024) It has become more important for social spaces to be secured, especially on the platforms such as Instagram to uphold user safety and trust. Detection of fake profiles which compromise the authenticity of these platforms are carried out with the help of machine learning techniques. Since machine learning models can analyse user behavior patterns, profile attributes and interaction data, they can correctly spot suspicious activities like bots, fake accounts and impersonators. Some of techniques which are used to identify inconsistency in profile, content, and engagement metrics are supervised learning, natural language processing and anomaly detection. For example, models can pick up on unusual activity patterns like rapid following/unfollowing, as well as generic image use, which are both a signal that the profile is fake. By constantly updating these models using large datasets, Platforms such as Instagram could achieve a safer and more realistic environment by minimizing spam, scams, and harassment alongside a true user experience.

T. Sudhakar et al(2022) This document ‘Fake Profile Identification Using Machine Learning’ aims to deploy advanced algorithms to detect and prevent the creation of fake accounts in social media platforms. Supervised and unsupervised machine learning techniques are applied on patterns in user behavior, profile characteristic, and interaction, which could be indicative of the presence of a fake profile. Trained on huge datasets comprising both genuine and fraudulent profiles, these systems are able learn to detect (what we call) anomalies like unnatural posting frequency, suspicious follower growth, or stock images being used. For even greater detection,

natural language processing (NLP) and image recognition help discern potential automation and manipulation by looking through text and visual content. As machine learning models continue to learn, they improve over time becoming better at identifying real users versus fake users. It protects online communities from spam, scamming and impersonation and creates a safer and more trustworthy social media environment.

Ananya Bhattacharya et al (2021) Machine learning techniques have now become important tools for the detection of fake profiles on social media in maintaining online integrity. Machine learning models can use user data to detect any patterns which indicate the fake account, like unusual activity, bot like behavior, the fake profile attribute etc. Supervised techniques are trained on labeled data that consists of both legitimate and fraudulent profiles and unsupervised learnings that detect new types of fraud without pre labeling. Textual data is analysed using natural language processing (NLP) to find irregularities like posts or comments that can be diagnosed as automated or fake behaviour. Profile pictures are fed into image recognition algorithms that look for sign of stock photos or manipulated images — other hallmarks of fake accounts. It also can find abnormal engagement patterns, such as rapid following or mass messaging. As these models improve, they keep misinformation, scams and impersonation spread lower, meaning online interactions are safer and more authentic.

Preethi Harris et al (2021) In my paper 'Fake Instagram Profile Identification and Classification using Machine Learning', we want to use machine learning to identify and classify with high accuracy the fake profiles on Instagram. Machine learning models are able to analyze features like profile metadata, user behavior, post frequency, user engagement patterns and detect anomalies that indicate fake accounts. Typically, supervised learning techniques such as decision trees or support vector machines are applied to classify profiles by taking into account datasets where some profiles are legitimate and some are fake. For assessing authenticity of captions, comments and direct message, natural language processing (NLP) and image recognition algorithms help determine signs of stock or stolen images in profile pictures. They also enable unsupervised learning techniques that can detect new patterns of fake profiles without having previously been labeled, so that the system can evolve against the changing patterns of fraud. Through this combination of methods, we achieve improved detection and classification, allowing Instagram users to have a safer and more authentic experience by diminishing scams, spam, and impersonation.

A. Bhattacharya et al (2021) However, as social media security became a top regard for the

public, the machine learning techniques came in handy to detect fake profiles on social media applications. Machine learning models examine user activity, profile data, and engagement patterns to detect unnatural frequency of posting, identifiable patterns of follow growth, and bot behavior, all of which are associated with widespread fake accounts. They are datasets that have been labeled and supervised learning algorithms are trained to differentiate between genuine and fake profiles using historical patterns. Examinations of textual data and detection of unnatural and repetitive language commonly used in bots mostly depends on Natural language processing (NLP). Stock photos, manipulated images or other signs of fake profile pictures are also identified using some image recognition techniques. Moreover, because they work with newly unknown behavior, methods for anomaly detection are capable of detecting a new type of fraudulent action from unseen patterns. Over time and as these models learn, they get better at spotting fake profiles and thus help curb scams, impersonation and distribution of misinformation on social media.

Kristo Radion Purba et al (2020) The paper's focus is based on the note and how one can classify fake users on Instagram by utilizing machine learning models. Generally, supervised learning algorithms are able to assess user behavior, profile data, as well as interaction patterns to be able to differentiate between genuine and fake accounts. They then train on labelled datasets of real and fake profiles to help train on how they'll know the difference between the two. In this method the common algorithms that are used are Decision Tree, Random Forest and Support Vector Machine (SVM) that are suitable for the classification task. For example, irregularities inside in measurable features like the number of posts per day, how posts are engaged with (i.e. how many likes, how many comments, how many followers' developers have), or textual features in captions and comments in conversations are detected. Similarly, image recognition is employed in them to catch stock photos or stolen images. Continuous training on newer and newer data helps with these algorithms to get better over time to combat the fake accounts, lower the spam and increase the secure level of the platform.

Padmaveni Krishnan et al (2020) This paper investigates the ability of finite automata (FAs) as a computational model in fake profile identification on online social networks, e.g., online social networking websites such as social media. Finite automata are mathematically model kind recognizing pattern by taking place of the input to the states and are applicable for abnormal behavior of fake accounts detection. A natural application of these automata is to train them to learn how to track a user's interaction such as frequency of pose, patterns of

follow/unfollow and a group of user generated engagement metrics which real users and artificially do not behave the same. Different states are associated by the finite automata to normal and suspicious behaviors, in order to keep record of user activity and to indicate anomalies indicative of a fake profile. These models are effective because they are efficient and they can handle large scale data as well as provide a real – time detection. However, as long as they sit in a social network, finite automata are a powerful too for help improve platform security and reduce the number of fraudulent accounts.

Kedir Lemma Arega et al (2020) The main title of this thesis is 'Social Media Fake Account Detection for Afan Oromo Language using Machine Learning'. Vernaculars of Afan Oromo have a special linguistic and cultural character that make traditional fake account detection methods unreliable. This approach would entail training of machine learning models to learn the patterns of user behavior, textual contents and patterns of engagement in Afan Oromo language. Finally, natural language processing (NLP) techniques are employed to analyse the posts and comments in Afan Oromo on structure, vocabulary and syntax in order to identify such abnormal usages that are a consequence of inconsistency or natural in language as to be found in automated bot (or fake accounts). To classify the user either as a real user or a fraud, supervised learning algorithms such as decision trees and support vector machines (SVM) are used on a dataset provided for specifying legitimate or fraudulent users. This approach aims as to enrich the social media security for native speakers of Afan Oromo language by tailoring models for the Afan Oromo language in order to improve the accuracy of fake account detection.

Methodology

The method of detecting fake profile with machine learning technique follows several systematic steps to collect data. In this step, user profile data is extracted from social media platforms such as number of friends/ followers, frequency of posting, content of posts, completeness of the profile, and account activity patterns that were used, among other attributes, the user ID and Public APIs can be used as a data source, as well as scraping the web (with adherence to compliance with privacy policies).

Then data preprocessing happens to provide clean, prepared data for analysis. This involves handling missing values, normalizing the numeric fields, encoding the categorical data, and dropping the irrelevant or duplicate features. Finally, feature selection is done to find the top

fakers' signs based on unusually high friend requests, repetitive comments, or low level of personal content.

Once the dataset is ready, the machine learning algorithms are ready to be utilized. The common techniques are supervised learning models namely Decision Trees, Random Forest, Support Vector Machines (SVM), and Logistic Regression, which are trained with labeled data (authentic vs. fake profiles). When labels aren't available, such as obtaining them, there are unsupervised methods like K-Means clustering, DBSCAN that can find anomalies or outlier behavior.

Metrics including accuracy, precision, recall and F1-score are used to evaluate the model after training to ensure effectiveness. Finally, the system is deployed in a pipeline for real time detection of new profiles as genuine or suspicious based on patterns learned so far.

This approach of fake profiles detection is a structured methodology that enables an accurate and scalable approach to this problem using machine learning.

Rationale of the study

Fake profiles on social media, tagged of fake accounts and fake profiles of visuals, paintings, compilations in forms of reading and other activities, have made the system more prone to the shock of fake profiles or accounts. These may be profiles created with the intention of misusing, effacing or destroying consumer or corporate confidence, or controlling public discourse, or that represent a considerable threat to personal or corporate security. Still, with the huge number of fake account creation, and advanced sophistication, traditional detection methods like manual moderation and rule-based systems are no longer adequate. Typically, they are static, time-consuming, and have a high error rate, rendering them inefficient for real-time, as well as large scale detection.

In this case, this rationale will be based on the fact that machine learning based methods (are) able to learn from data, adapt to new patterns and make intelligent decisions with little to no human intervention. There is no doubt that machine learning approaches have demonstrated great success in identifying complicated, nonobvious behavioral pattern exhibited by fake profiles. The goal of this study is to inspect, compare, and assess the abilities of different machine learning techniques to distinguish between genuine and fake profiles on the context of different online platforms.

The study synthesizes recent accomplishments, while bringing up strengths, limitations, and possible improvements to contribute to development of more secure, scalable and smarter

detection systems, allowing more security and reliability of online environments for genuine users.

Conclusion

In the advancement of safe and trusted online platform, machine learning approaches in fake profile detection have been made possible. Traditional detection methods for fake accounts have demonstrated weaknesses in dealing with their aptitude for dynamism and adaptability, as fake accounts get increasingly complex and scalable. Leveraging huge amounts of user data, machine learning can more robustly intelligently and more scalable discover subtle patterns and behaviors of fake profiles. Supervised learning (e.g Random Forest, SVM), unsupervised learning (e.g clustering, anomaly detection) and deep learning models by various researchers proved to adequately classify and predict fraudulent activities. In addition, the use of natural language processing, social graph analysis, and hybrid model has enhanced the detection accuracy and response times even further. Despite that, there are still challenges which are still need to be resolved, for example, creating high quality labeled datasets, handling class imbalance, and overcoming adversarial techniques used by smart attacker. Nonetheless, constant research and technological innovations are steadily enhancing the matter of machine learning systems in this domain. With more ubiquitous pressure to address issues of user security and regulatory compliance, the urgency of implementation for intelligent detection systems on platforms require immediate attention. On the whole, fake profile detection based on machine learning contributes to more integrity to the platform and maintains a more secure digital environment by preventing the creation of other malicious and deceptive profiles. Partly due to this review, these techniques are urged to be continually improved and adjusted in the aim of gaining the upper hand in the fight against all forms of online fraud and abuse.

Reference

1) Goyal, B., Gill, N.S. & Gulia, P. Securing social spaces: machine learning techniques for fake profile detection on instagram. Soc. Netw. Anal. Min. **14**, 231 (2024).
2) T. Sudhakar, B. C. Gogineni and J. Vijaya, "Fake Profile Identification Using Machine Learning," 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Naya Raipur, India, 2022, pp. 47-52
3) B. Prabhu Kavin et al., "Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks", Wireless Communications

and Mobile Computing 2022, 2022.

4) Ananya Bhattacharya, Ruchika Bathla, Ajay Rana and Ginni Arora, "Application of Machine Learning Techniques in Detecting Fake Profiles on Social Media", the 9th International Conference on Reliability Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Sep 3-4, 2021.

5) Preethi Harris, J Gojal, R Chitra and S. Anithra, "Fake Instagram Profile Identification and Classification using Machine Learning", 2021 2nd Global Conference for Advancement in Technology (GCAT), Oct 1-3, 2021.

6) A. Bhattacharya, R. Bathla, A. Rana and G. Arora, "Application of Machine Learning Techniques in Detecting Fake Profiles on Social Media," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-8

7) Sreeram Gutha, K Srinivasa Rao and B. Deevena Raju, "Detecting Fake Account on Social Media using Machine Learning Algorithms", International Journal of Control and Automation, vol. 13, no. 1s, pp. 95-100, 2020.

8) Kristo Radion Purba, David Asirvatham and Raja Kumar Murugesan, "Classification of instagram fake users using supervised machine learning algorithms", International Journal of Electrical and Computer Engineering, vol. 10, no. 3, pp. 2763, 2020.

9) Padmaveni Krishnan, D. John Aravindhar, Palagati Bhanu and Prakash Reddy, "Finite Automata for Fake Profile Identification in Online Social Networks", Proc. Of ICICCS 2020.

10) Kedir Lemma Arega, "Social Media Fake Account Detection for Afan Oromo Language using Machine Learning", New Media and Mass Communication ISSN 2224–3267 (Paper) ISSN 2224–3275 (Online), vol. 90, no. 2020.

11) Samala Durga Reddy, "Fake Profile Identification using Machine Learning", IRJET journal, vol. 06, no. 12, pp. 1145-1150, Dec 2019.

12) Mudasir Ahmad wani, Nancy Agarwal, Suraiya Jabin and Syed Zeeshan Hussain, "Analyzing Real and Fake users in Facebook Network based on Emotions", the proceedings of the 2019 11th International Conference on Communication Systems Networks (COMSNETS), pp. 110-117.