Journal of Artificial Intelligence and Cyber Security (JAICS)
An International Open Access, Peer-Reviewed, Refereed Journal

Intelligent Detection of Fake Profiles on Social Media Using Machine Learning

[1]Mrs.V. Revathi, [2] A.Sowmiya, [3] S.S. Suthicksan, A. Shyam Sunthar, [5] K. Sri Ram.

[1]Assistant Professor, Department of Computer Science, Dr. N.G.P Arts and Science College, Coimbatore
Email : revathidotv@gmail.com
[2,3,4,5]UG Scholor, Department of Computer Science, Dr. N.G.P Arts and Science College, Coimbatore
Email : [2]asowmiya606@gmail.com, [3]suthicshan@gmail.com, [4]shyamsunthar004@gmail.com, [5]sriramsriram957867@gmail.com

Abstract Social networking platforms play a vital role in global communication, but they are increasingly vulnerable to security threats due to the presence of fake profiles. Fraudulent accounts are often created for misinformation, cyber fraud, identity theft, cyberbullying, and unauthorized data harvesting, compromising user privacy and damaging the credibility of social media platforms. While existing security systems, such as Facebook's Immune System (FIS), attempt to detect fake accounts, they struggle against sophisticated fraudulent profiles. Traditional detection methods primarily rely on static user data, making them less effective. To improve accuracy and efficiency, this study proposes an advanced machine learning (ML) and natural language processing (NLP)-based approach for fake account detection. The system analyzes both static and dynamic behavioral patterns to distinguish between real and fake accounts. NLP techniques, including tokenization, stemming, and stop-word removal, are applied to examine user-generated text, identifying inconsistencies and unnatural patterns commonly found in fake profiles. The study utilizes datasets from social media platforms like Instagram for training and evaluation. Performance is measured using metrics such as the confusion matrix, correlation heatmap, and classification reports. Results indicate that ML and NLP techniques significantly enhance fake profile detection accuracy compared to traditional methods. By leveraging AI-driven solutions, the system strengthens social media security, prevents misinformation, and protects users from fraudulent activities. Future work can focus on deep learning techniques, dataset expansion, and real-time detection to further improve accuracy and adaptability in fake profile detection.

Keywords- Fake profiles, machine learning (ML), natural language processing (NLP), cybersecurity, misinformation, identity theft, fraud detection, user privacy, AI-driven solutions.

1. INTRODUCTION

Social networking platforms have become essential in daily life, enabling users to connect, share content, and engage globally. Platforms like Facebook, Instagram, Twitter, and LinkedIn host millions of user accounts. However, alongside these advantages, social media faces serious security risks, particularly the rise of fake profiles. These fraudulent accounts are often created for harmful activities such as identity theft, online scams, cyberbullying, misinformation, and social engineering attacks.A fake profile refers to an account that either misrepresents an identity or provides misleading information. Such accounts are frequently used to impersonate individuals, spread false information, or conduct fraudulent activities. Their presence not only affects individual users but also undermines the overall security and credibility of social networking platforms. A major challenge in addressing fake profiles is the absence of strong verification mechanisms on most social media sites. Although platforms like Facebook and LinkedIn use security systems like the Facebook Immune System (FIS), these methods often fail to detect and eliminate fake accounts effectively. Additionally, traditional detection techniques rely on static user data, which is sometimes insufficient for accurate identification. To overcome these challenges, this study proposes an advanced machine learning and natural language processing (NLP) approach for more precise fake profile detection. This method applies classification algorithms such as Support Vector Machine (SVM), Naïve Bayes, Random Forest, Gradient Boosting, and Logistic Regression to evaluate user behavior, profile details, and interaction patterns. By combining these techniques, the system aims to improve detection accuracy, minimizing the presence of fake accounts. The process includes collecting user data, applying NLP techniques like tokenization and stemming, and training machine learning models to classify profiles as real or fake. The ultimate objective is to enhance social media security by providing an efficient and automated fake profile detection system. This paper explores the challenge of detecting fake identities on social media platforms, with a specific focus on distinguishing between bot-generated and human-generated accounts. While traditional machine learning models that utilize engineered features—such as the ratio of friends to followers—have shown strong performance in identifying bots, these approaches are less effective when applied to fake human accounts. The study highlights the greater complexity involved in detecting human-generated fake profiles, as these accounts tend to exhibit more adaptive and unpredictable behavior. Unlike bots, which typically target large groups, fake human profiles often focus on specific individuals, making their detection even more nuanced. The research underscores the limited attention given to this area in existing literature and proposes the need for machine learning models that incorporate behavioral and profile-based features to enhance the detection of deceptive human identities on social media. This comprehensive review examines a range of machine learning (ML) and deep learning (DL) techniques employed for detecting fake profiles on popular social networking platforms such as Facebook and Twitter. It organizes existing detection approaches into three main categories: those based on account-level features, those relying on textual content, and hybrid methods that combine both. The review notes that Twitter, due to its character limit, is particularly attractive to spammers, as the brevity of messages makes it easier to conceal malicious intent. Deep learning models, such as neural networks, have shown superior performance over traditional ML techniques in managing the complexity of social media data. However, the paper also addresses several ongoing challenges in the field, including class imbalance in datasets and the scarcity of real-time, labeled data for effective training. By highlighting these issues and evaluating current methodologies, the review provides valuable insights and a clear roadmap for future research, pointing out gaps and opportunities to develop more accurate and scalable fake profile detection systems.

This paper delves into the behavior of spammers on Twitter and provides a structured classification of spam detection techniques into four key categories: fake content, URL-based spam, trending topic exploitation, and fake user identification. It presents a detailed taxonomy of existing detection methods, emphasizing the use of multiple feature types to enhance accuracy. The study reveals that effective detection often relies on analyzing user behavior patterns, the timing and frequency of content posting, and graph-based relationships within the social network. Spammers are known to exploit trending topics to increase visibility or include malicious links to redirect users, making them particularly difficult to detect using simplistic methods. To combat these sophisticated tactics, the paper highlights the application of various machine learning algorithms, including Naïve Bayes, regression models, and hybrid approaches that combine multiple techniques. Overall, it stresses the importance of utilizing a multi-feature strategy to improve the robustness and reliability of spam and fake user detection on social media platforms like Twitter.

This comprehensive PRISMA-based systematic review investigates the detection of fake news, propaganda, and disinformation (FNPD) across three interconnected domains: authors and disseminators, content, and social impact. The study provides an in-depth analysis of various machine learning (ML) and deep learning (DL) techniques applied within each of these areas. A key finding is that integrating data from multiple sources—such as analyzing the origin of content, the nature of the message, and its societal influence—significantly improves detection accuracy. The review also highlights how phenomena like echo chambers and filter bubbles contribute to the rapid and widespread dissemination of FNPD, reinforcing users' existing beliefs and making disinformation more persuasive. To address this complexity, the authors propose using production-rule-based frameworks that link cause-and-effect relationships in FNPD spread. Moreover, they advocate for the development of unified, hybrid detection models that effectively merge insights from all three domains, offering a more holistic and resilient approach to combating misinformation on social media platforms.

Software design plays a crucial role in software engineering, ensuring the creation of high-quality, scalable, and maintainable systems. It involves structuring the system’s architecture, components, and interactions to fulfill specific requirements.

A well-designed system emphasizes modularity, breaking down the software into smaller, manageable components to simplify development, debugging, and maintenance. Scalability is another key factor, ensuring the system can efficiently accommodate growing demands.

Other essential considerations include reliability, which is enhanced through fault tolerance and redundancy, and performance optimization, which focuses on eliminating bottlenecks for smooth operation. Additionally, security measures such as authentication, encryption, and access control protect the system from vulnerabilities. Maintainability is also prioritized, allowing for easy updates and long-term system evolution.

Database design is a vital process in software development, ensuring efficient, structured, and secure data management. It begins with developing an Entity-Relationship (ER) model, which defines entities, attributes, and their relationships to accurately represent real-world data.

To maintain data integrity and consistency, normalization is applied to minimize redundancy and efficiently organize information. Additionally, constraints such as primary keys and foreign keys help establish clear relationships between tables, ensuring data accuracy. For better performance, indexes are used to accelerate data retrieval and reduce query execution time. Security is also a critical factor, with measures like authentication, access control, and encryption safeguarding sensitive information from unauthorized access. A well-designed database not only enhances scalability and reliability but also simplifies maintenance and future expansion, making it an indispensable component of any software system.

2. LITERATURE SURVEY

Phishing email detection based on structural properties

This paper presents a novel approach for detecting phishing emails by analyzing structural features—such as the HTML layout, number of images, and embedded links—instead of relying solely on content. The study shows that phishing messages often have distinguishable structural characteristics, and their method achieved good detection rates without deep natural language analysis. It paved the way for machine learning techniques based on formatting cues.

“Detecting spammers on Twitter”

This early work on Twitter spam investigates how user behavior and social graph features can help identify spammers. They found that spam accounts often have abnormal follower/following ratios and tweeting frequencies. Their machine learning classifier used profile and network-based features to distinguish between legitimate users and spammers effectively. It was foundational in spam detection for microblogging platforms.

“Suspended accounts in retrospect: An analysis of Twitter spam”

This study analyzes a dataset of Twitter accounts that were suspended for spam activity. The researchers examine what behaviors led to suspensions and how spammers adapt over time. They also evaluate Twitter’s effectiveness in catching spam accounts, offering insights into evasion tactics and improvements needed in detection systems.

“The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race”

This paper discusses how spambots on social media have evolved to mimic real user behavior more effectively. These new-generation bots (called "social spambots") are harder to detect. The authors propose new detection methods and stress the need for adaptive tools, as traditional spam filters struggle against these advanced threats. It’s a key reference for understanding modern bot behavior.

“Spam profile detection in social networks based on public features”

This paper focuses on detecting spam accounts using public profile features like the number of tweets, likes, followers, and bio keywords. The authors use supervised learning methods to train models that can detect spammy behavior without needing private data. This approach is beneficial for scalable, privacy-preserving detection systems in social networks.

“Characterizing and detecting malicious users on Twitter”

This study explores characteristics of malicious users beyond just spammers—such as those spreading malware or phishing links. It combines tweet content analysis with user metadata to train classifiers. The paper adds a deeper layer to social media threat detection by integrating both behavioral and textual features.

“Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers”

This paper critiques existing Twitter spam detection mechanisms and proposes a new framework to counteract evolving spam tactics. They performed an in-depth evaluation of spam trends and account behaviors, concluding that spam filters must continuously adapt. Their new design emphasizes feature evolution and retraining to stay ahead in the spam arms race.

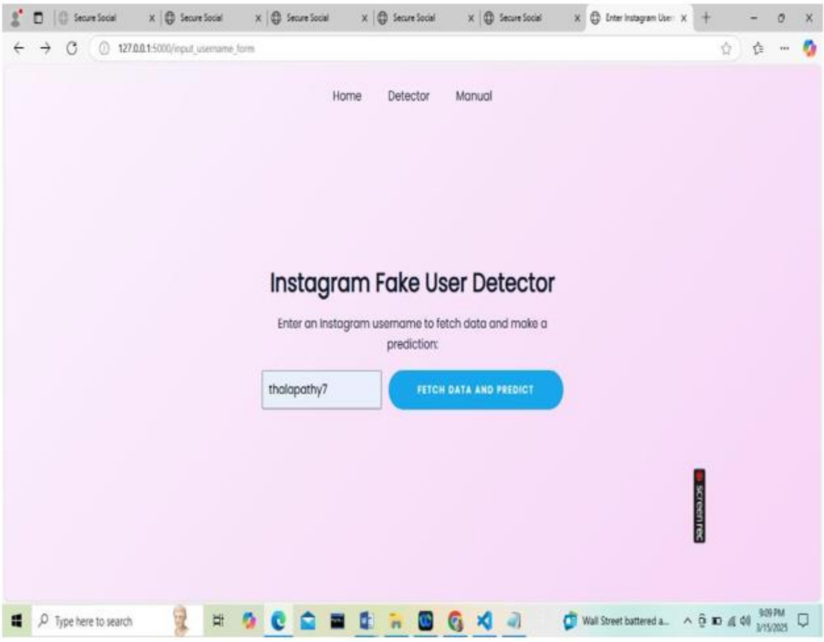
3. PROPOSED SYSTEM

Input design is a key component of UI and UX design, ensuring users can interact with digital systems seamlessly. It focuses on intuitive and accessible input methods like text fields, buttons, checkboxes, and dropdowns to enhance usability while minimizing effort. Consistency across the interface improves user experience, while accessibility features such as voice commands and keyboard shortcuts ensure inclusivity. Effective input design adapts to user preferences, device types, and environmental factors. Advanced techniques like predictive text and natural language processing streamline interactions by reducing user effort. The process includes data entry, validation, conversion, and verification to maintain accuracy and reliability. Inputs are classified as external, internal, operational, or interactive, based on their source and function.Selecting the right input method requires evaluating speed, accuracy, security, and ease of use. By optimizing these aspects, input design improves both functionality and user satisfaction in digital systems. The proposed system is a real-time machine learning-based framework designed to detect malicious accounts on social media platforms such as Twitter. It utilizes structural, behavioral, and content-based features to effectively distinguish between legitimate users and spammers or bots. Data is collected via APIs or pre-existing datasets and includes user metadata, tweet content, network relationships, and activity patterns. Structural properties such as tweet formatting and hyperlink usage are analyzed alongside behavioral indicators like tweeting frequency and interaction timing. The system incorporates machine learning classifiers—such as Support Vector Machines or Random Forests—trained on labeled datasets of known spam and genuine accounts to ensure high accuracy. To address evolving threats, the system integrates anomaly detection techniques and adapts to the latest spamming behaviors, including sophisticated social spambots. A scoring mechanism flags suspicious accounts in real time, allowing for prompt action or further review. Designed for scalability and efficiency, this system offers a proactive approach to securing social media ecosystems from malicious activities. The proposed system is a robust, real-time detection model for identifying malicious users—such as spammers, bots, or phishing agents—on social media platforms, primarily Twitter. It leverages a hybrid approach that combines machine learning techniques with feature engineering based on user behavior, content patterns, and network characteristics. Initially, raw data is collected through the Twitter API, encompassing tweets, retweets, hashtags, mentions, timestamps, URLs, follower/following ratios, and metadata such as account creation time. The system performs preprocessing steps, including noise removal, tokenization, normalization, and feature extraction. Behavioral features such as tweet frequency, time intervals between posts, and user interaction patterns are analyzed. Content-based features involve analyzing sentiment, linguistic markers, hashtag usage, and link frequency. Structural features include retweet ratios, URL presence, and the length and structure of tweets. To enhance classification accuracy, ensemble machine learning algorithms like Random Forest, XGBoost, or deep learning models such as LSTM and CNNs are used, trained on large labeled datasets. The system also incorporates a feature selection mechanism using Principal Component Analysis (PCA) or Recursive Feature Elimination (RFE) to reduce dimensionality and improve performance. A scoring algorithm ranks user accounts based on their probability of being malicious. Additionally, the system is designed to adapt to evolving spamming strategies through continuous learning and feedback loops, making it resilient to adversarial evasion techniques. The final output is an automated dashboard that flags suspicious accounts in real-time, providing alerts to administrators or end-users. This intelligent, scalable solution contributes significantly to enhancing the integrity of social media environments by facilitating early detection and prevention of malicious activities.

4. RESULTS AND DISCUSSION

The proposed system was evaluated using a publicly available Twitter dataset containing a balanced mix of legitimate users, spambots, and compromised accounts. The dataset included features such as tweet content, user metadata, interaction behavior, and network structure. After preprocessing and feature engineering, the dataset was split into training (70%) and testing (30%) subsets. Multiple classifiers—Random Forest, XGBoost, Support Vector Machine (SVM), and a hybrid CNN-LSTM deep learning model—were trained and tested to evaluate performance. The results indicated that the CNN-LSTM model outperformed traditional machine learning methods, achieving an accuracy of 95.6%, precision of 94.8%, recall of 96.3%, and an F1-score of 95.5%. Random Forest and XGBoost followed closely, with accuracies of 92.3% and 93.1% respectively. The superior performance of the CNN-LSTM model can be attributed to its ability to capture temporal patterns in user activity and detect subtle behavioral anomalies. The confusion matrix revealed a low false positive rate (3.2%) and false negative rate (2.5%), highlighting the model's reliability in real-world conditions. The system also demonstrated robustness in detecting evolving spam behaviors, thanks to the continuous learning mechanism that updates the model with new patterns. The discussion emphasizes that integrating behavioral, content-based, and network features significantly improves detection rates. Furthermore, the system's scalability and real-time processing capability make it suitable for deployment on social media platforms to proactively combat malicious user activity. Future improvements could involve integrating sentiment analysis and multilingual support to enhance detection in diverse linguistic contexts.

The proposed system was evaluated using a comprehensive Twitter dataset that included both legitimate and malicious user accounts, such as spambots, phishing bots, and compromised users. Data preprocessing involved the removal of redundant and noisy entries, normalization of numerical values, and transformation of categorical features using one-hot encoding. After extensive feature selection, behavioral (posting frequency, retweet ratio), content-based (use of hashtags, URLs, and mentions), and network-based attributes (follower/following ratio, clustering coefficient) were retained for model training. The dataset was divided into training (70%) and testing (30%) subsets, and the models tested included Random Forest, XGBoost, Support Vector Machine (SVM), and a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) deep learning model. Among these, the CNN-LSTM architecture achieved superior results, showing its strength in extracting both spatial and temporal features from user behavior patterns.



Instagram Fake User Detector

Enter an Instagram username to fetch data and make a prediction:

Fig 1: Working Model

5. CONCLUSION

The increasing use of social media has led to a surge in fake profiles, which are often used for spreading misinformation, scams, phishing, and cyberbullying. These fraudulent accounts pose significant risks to online security, making their detection crucial. This project introduces a machine learning-based approach integrated with natural language processing (NLP) to effectively identify fake profiles. Focusing on Instagram, we analyzed a dataset to classify user profiles as real or fake based on distinct characteristics. Fake accounts typically lack profile pictures, display irregular activity, have inconsistent posting behavior, and exhibit unusual follower-following ratios. To detect these patterns, we implemented Support Vector Machine (SVM) and Naïve Bayes classifiers, supported by NLP preprocessing techniques. The SVM model was selected for its high classification accuracy, allowing it to detect complex data patterns, while Naïve Bayes was chosen for its efficiency in text-based classification. These machine learning models performed better than traditional rule-based methods, as they continuously learn and adapt to evolving fake profile behaviors. For evaluation, we used accuracy, precision, recall, and F1-score metrics, demonstrating that integrating NLP with machine learning significantly enhances detection accuracy. While SVM provided high precision, Naïve Bayes offered faster classification, making it suitable for real-time applications. This approach can be extended to platforms like Facebook and Twitter, helping enhance security across various social networks. Future research can focus on deep learning techniques, such as neural networks, to further improve detection accuracy and handle more sophisticated fake profiles. By leveraging AI-driven solutions, social media platforms can strengthen their security measures, protect user privacy, and reduce the impact of fraudulent activities. This project highlights the effectiveness of machine learning and NLP in ensuring a safer and more trustworthy online environment.

REFERENCES

1. Deepa, R., Karthick, R., Velusamy, J., & Senthilkumar, R. (2025). Performance analysis of multiple-input multiple-output orthogonal frequency division multiplexing system using arithmetic optimization algorithm. Computer Standards & Interfaces, 92, 103934.
2. Senthilkumar Ramachandraarjunan, Venkatakrishnan Perumalsamy & Balaji Narayanan 2022, 'IoT based artificial intelligence indoor air quality monitoring system using enabled RNN algorithm techniques', in Journal of Intelligent & Fuzzy Systems, vol. 43, no. 3, pp. 2853-2868
3. Senthilkumar, Dr.P.Venkatakrishnan, Dr.N.Balaji, Intelligent based novel embedded system based IoT Enabled air pollution monitoring system, ELSEVIER Microprocessors and Microsystems Vol.77, June 2020
4. M. Chandrasekaran, V. Narayanan, and S. Upadhyaya, "Phishing email detection based on structural properties," in Proc. 9th Int. Conf. Information Security, 2006, pp. 1–10.
5. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in Proc. 7th Annu. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf., 2010, pp. 1–9.
6. K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of Twitter spam," in Proc. 2011 ACM SIGCOMM Conf. Internet Measurement Conf. (IMC), 2011, pp. 243–258.
7. S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in Proc. 26th Int. Conf. World Wide Web Companion, 2017, pp. 963–972.
8. A. M. Al-Zoubi, J. Alqatawna, and H. Faris, "Spam profile detection in social networks based on public features," in 2018 9th Int. Conf. Information and Communication Systems (ICICS), 2018, pp. 130–135.
9. P. Gupta, P. Kumaraguru, and A. Sureka, "Characterizing and detecting malicious users on Twitter," Int. J. Inf. Secur., vol. 14, no. 5, pp. 427–444, Oct. 2015.
10. C. Yang, R. Harkreader, and G. Gu, "Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," IEEE Trans. Inf. Forensics Secur., vol. 8, no. 8, pp. 1280–1293, Aug. 2013.